UNIVERSITETI I EVROPËS LIGLINDORE
УНИВЕРЗИТЕТ НА ЈУГОИСТОЧНА ЕВРОПА
SOUTH EAST EUROPEAN UNIVERSITY

FACULTY OF CONTEMPORARY SCIENCES AND TECHNOLOGY

PHD DISSERTATION

TOPIC: "A MODEL FOR AUTOMATED MATCHING BETWEEN JOB MARKET DEMAND AND UNIVERSITY CURRICULA OFFER"

Supervisor:

Assoc. Prof. Dr. Artan Luma

Candidate:

MSc. Ylber Januzaj

Tetovo, 2019

Table of Contents

List of figures

List of tables

Abstract

Nowadays technology has reached a very high point of development and research, but it can still be considered that it has remained to be discovered. The application of technology is of great importance in almost all fields, directly affecting their development and advancement. However, technology as a field remains a challenge for graduate students but also for different companies as it is not possible to fill the vacancies in this field from the same student. The reason why these vacancies are not being met in the field of technology lies precisely in the fact that there is a discrepancy between the curricula offered by the Universities and the labor market demands currently in existence. In this regard, we have exploited technology precisely to fill this space, directly contributing to the creation of an automated model that will be able to compare between university curricula and labor market demands. The creation of this automated model will be done by using different techniques that will give us accurate and quick results. These techniques come from the Data Mining field as well developed and advanced fields.

Numerous analyzes have been made in this field, trying to make different comparisons and adjustments between university curricula and labor market demands, but it is not yet possible to create a model that will make automated comparison.

As we have already mentioned above, the aim of this topic is to create an automated model that will be able to compare the demands of the labor market and curricula currently offered by universities in the field of technology. The first step is the dissemination of questionnaires in private and public universities in Kosovo and Macedonia, in order to obtain samples for different views regarding the teaching of these Universities. These results we have managed to get from the teacher and the students have served to answer some of the scientific questions and hypotheses that we raised earlier. The second step is experimentation, where through Web crawling techniques we will be able to get information on open competitions from different companies that are published on different websites. Then, using the Data

Mining techniques, we will make pooling of university competitions and curricula so that we can make a comparison between these profiles.

The creation of such a model is a major contributor to higher education, as it will provide us with the exactitude of the fit between the labor market demands and the curricula offered by universities. Beneficiaries in this case will not only be universities, but will also be different private companies as through this model will be able to meet the gaps that are currently in the field of technology for newly graduated students. And most importantly, besides universities and private companies, such a model will bring great benefits to Triple Helix as a very important system between universities, government and private companies.

1. Introduction

Technology plays a very important role in virtually all areas, and has become an inseparable part of the industry. Currently, industry and technology are at a high point of development and research, but there is an ever increasing gap between the market needs and the skills that universities deliver to students. There is an increasing need for consolidation between university curricula and the industry needs in terms of qualifications.

Studying relevant program has a high importance for students and the labor market. Currently in the field of technology, there are many labor market demands that fail to be met for various reasons, one of which is the lack of professional preparation in the relevant field. Applying technology to the industry also has a high importance, as it directly affects the country's income, almost in all areas. In (Anicic, Divjak and Arbanas, 2017) we can observe that the importance of technology lies on the fact that 5% of Europe's Gross Domestic Product (GDP) is directly dependent on the field of technology. How technology is important, is shown in some research that has been done on trends and job requirements in this area for the period 2012-2020. In (Anicic, Divjak and Arbanas, 2017) it is mentioned that, according to the *"Main Forecast Scenario"* from 2012, an increase of 7.4 million to 7.9 million job vacancies is expected in 2020, according to *"Stagnation Scenario"*, an increase of up to 7.8 million job vacancies by 2020, and according to *"Disruptive Boost"*, an increase of 8.1 million job vacancies is expected by 2020. Therefore, based on these forecasts in all three cases, which are claimed to be optimistic values, job vacancies in the field of Technology will exceed the number of graduates in this area, and at the same time we can notice the great importance of studying relevant programs in the field of Technology.

Earlier, it was mentioned that there are many reasons that impact the requirements to be not sufficiently met by graduate students in the market, such as lack of experience, poor knowledge of international languages, and weaknesses in communication, as well as studying in the right direction based on labor market requirements. In (Aziz & Yusof, 2016) it has been mentioned that as far as it is important to complete the studies and get the status as a "graduate", it is also important to determine the quality of the program offered by the

relevant institution. Based on the importance of such researches, the application of some sub-fields of technology is required in order to have a more accurate and optimistic result. Such research also requires a combination of the field of technology and mathematics, as besides Data Mining methods, such as classification, clustering, etc., also mathematical techniques such as Bayes, Regression, Decision Tree, descriptive statistics, etc. will be applied. Below we will present the latest achievements in this regard, as well as the recommendations that researchers provide for the future.

Currently there is a very large amount of data in the information industry. But these data, if not converted into useful information, remain as unnecessary data. It is imperative that these data be converted and extracted from the information they can use. It is not necessary to do only extract of these data, but we need also to do data cleaning, data integration, data transformation, data mining, and data presentation.

So, Data Mining can be defined as the process of extracting knowledge from the data. The knowledge that can be extracted and used in several areas such as:

- **Market Analysis.**
- **Education.**
- **Customer Retention.**
- **Production Control.**
- **Medical Analysis.**

So, as we can see, Data Mining can also be applied in the Education field. In our case we use it for classification and data clustering, and with the application of Data Mining we ensure that we have very effective and reliable classification and clustering.

On the other hand, Machine Learning is a field that has enabled us to automate many processes that we use today to ease our work in many aspects. In our case we will apply Machine Learning in order to automate our system to be able to give recommendations.

1.1. Problem

The increment of graduate students which are unable to find themselves in labor market due to inadequate education is the main problem that makes us create such a model. As noted above, despite the fact that the number of students is increasing, the labor market demand in the field of technology is increasing dramatically every day.

There are currently 30 institutions of higher education in Kosovo, out of which 9 are public universities and 21 are private colleges. All these institutions of higher education offer totally 468 study programs at the three levels of study: bachelor, master and phd. As our focus is in the field of technology, according to the research that is made in the data published by KAA[1], there are 20 faculties that offer technology studies. As far as the number of studies in the field of technology offered by these faculties is 68, with 38 being bachelor's, 26-level master's and 4-level doctoral degrees.

As part of our research we have also been the survey of private companies that offer work in the field of technology. Based on the analysis of the results we received from these companies, we have noticed that employers are always looking for students who are prepared in different fields: programming, databases, computer networks, etc. According to employers, even after employing students in their companies, they are obliged to keep these students' practices and to send them to different trainings so that they can prepare for the workplace in the best possible way and to be more professional.

The other problem is that universities in Kosovo should not graduate as students who are only prepared for the local or regional labor market, as students must be prepared for the European market as well. According to statistics, labor market demands in the field of technology are always increasing, and as mentioned above, in 2020 these job requirements will exceed the number of graduates in the field of technology. Therefore, by creating our

[1] http://www.akreditimi-ks.org/new/index.php/en/

model, we will manage to solve this problem by always offering students who are also prepared for the European market.

1.2. Motivation

It is precisely the data in previous that have motivated us to do such a research on the creation of an automated model, since even though there are a considerable number of study programs there are still positions that cannot be met by graduate students in the field of technology. The reasons why these positions cannot be met are different, ranging from: communication skills, foreign language knowledge, appropriate field practice, participation in mobility, etc.

The number of positions in the field of technology in the future is exactly what motivated us to make such a model. If we analyze the high number of study programs that are accredited then we can easily conclude that there is a non-compliance with the labor market requirements. We argue that there is a discrepancy between university curricula and labor market demands because currently there are positions that are open in the field of technology but lack the staff to fill those positions.

It is also the development and advancement of techniques we will apply what motivated us to create such an automated model.
Currently, all announcements that are currently announced regarding free positions that occur in the field of technology are made through web pages, emails and newsletters. It is important for us to determine what is the most widespread form of how students are informed about open competitions in the field of technology. Based on the questionnaires that have been disseminated to the students and the students have identified the most widespread form of how students are informed about open competitions in the field of technology through web pages. Also this has been a motivation for us to apply web crawling and scraping techniques in order to get all the information from websites that publish open competitions in the field of technology. Given that the actual webpages besides the information are also focused on the design content, it has been the web crawling method

which has enabled us to make accurate information extracted from these web pages. Based on the results of a website since we have applied web crawling techniques, we have dozens of free open positions in the field of technology that in their description require different knowledge in .Net, Java, Android and IOS Developer.

Also what has motivated us to create such a model is the fact that public universities have published all the syllabuses that they offer on their websites. These syllabuses will also be taken and processed in order to make it possible to apply Data Mining collection methods, thus creating a variety of syllabus profiles offered by universities.

Currently, universities in our country are working hard to make a fit between their curricula and labor market demands. But based on the results surveyed by different universities, none of these universities have accurate statistics on labor market demands. Also this is a motivation for us because through such research we will be able to present accurate results on the labor market demands in Kosovo and Macedonia. The ways in which universities are trying to make adjustments between labor market researchers and their curricula are diverse, ranging from the establishment of industrial boards to the organization of workshops and conferences in this field.

All of these reasons are mentioned which have motivated us to create such a model that will accurately reflect the level of the actual adaptation of labor market demands and curricula offered by universities. Based on these results we will be able to provide feedback on changing curricula so that they fit the market to a higher level.

1.3. Objectives

Our main idea is ***"Automated model for comparison of labor market and University curricula's".*** While this is our target we also emphasize the importance of the research:

1. **Universities need to adjust their Curricula to labor market.**
2. **Labor market demand profile "overflow".**

Now in one side we have the main idea, and in other side we have the importance of the research. Based on this now we are able to define the problem which is known as the main gap of the research:

Are we able to automate the process of profile – curricula matches?

The idea is to make an automated comparison between market demands and university study programs by applying Data Mining techniques. First we scrap information from job vacancies web sites. How we do this? Actually there are web sites which publish job vacancies in different fields, our focus will be in the field of technology. The Information that will be extracted are:

- **Title of the vacancy.**
- **Position.**
- **Skills that are required.**
- **Advantages.**

These information will be processed and will be clustered in order to create different types of profiles.

In other side we have the information about University curricula that will be extracted. While Universities in our region allow public access on study programs, these information will be automatically extracted. The information that will be extracted are:

- **Title of study program.**
- **List of subjects M or E.**
- **Credits of subjects.**
- **Number of hours for Lecture/Exercises.**
- **Description.**
- **Career.**
- **Learning Outcomes.**

Also these information's will be processed and will be clustered in order to create different types of profiles. After we have groups of profiles in both sides, we apply matching algorithm in order to compare the adjustment between curricula's to labor market. Next in Figure 3 we show a sketch of our automated model.

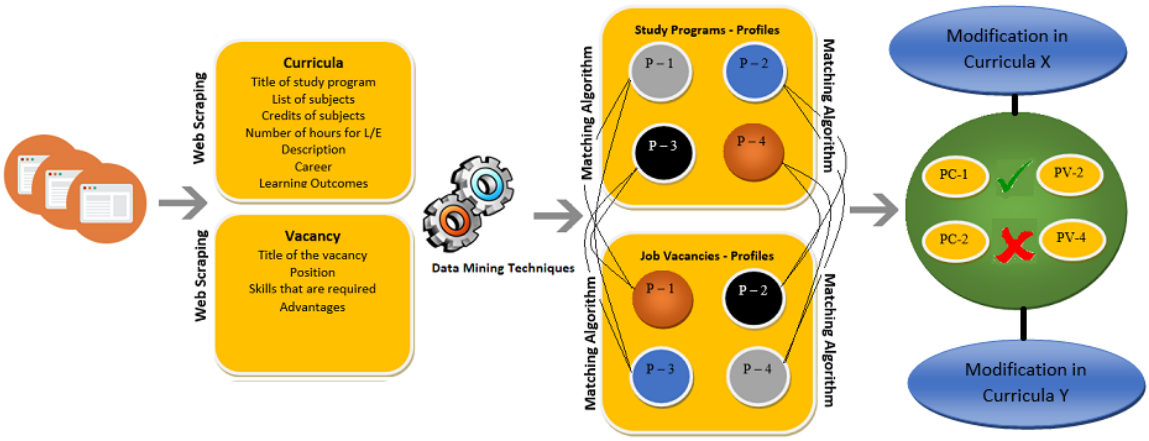


Figure 1. Sketch of automated model

In Figure 1 we can see the sketch of our proposed automated model for profile – curricula matching. First we have the scraping process from web sites, here we extract the content for the curricula and for vacancies. After we extract the data, now we apply Data Mining Techniques and we get different groups of profiles. After we have profiles now then we apply matching algorithms and we get matching between Profile Curricula "PC" and Profile Vacancies "PV". Finally after we get the matching between PC and PV, our model is able to give recommendation for any kind of modification in Curricula X or Y. These kind of modification includes removing a subject or adding a new one.

1.4. Hypotheses

1. **Automated methods can help matching curricula with workforce demand.**
2. **Curricula offer and job demands can be "quantified" using Data Mining Techniques.**
3. **Proposed model can provide suggestions for curricula improvements.**

1.5. Research Questions

1. **What are the requests and recommendations in cooperation between market and universities?**
2. **Will automated profile – curricula matching decrease the adjusting time of companies?**
3. **How Universities are able to adapt to the recommendations that automated model exposes.**
4. **What is the current cooperation between universities – market in order to improve the quality of skills and their match with labor market demands?**

1.6. Proposed Model

- From specific web sites extract specific content using special tools for scraping. In this phase we have to identify the web sites that publish vacancies in the field of the technology. After we identify the links, now we have to identify the format of the text that the vacancies are published.
- Process extracted data and apply Data Mining Techniques for clustering in order to have a high dimensional data clustering. We use Clustering while we need to create groups of different profiles which are based on the description of each vacancy and each curricula. These data will be used later for matching.
- Using the previous data which are separated in groups, now we apply matching algorithm in order to make matches between two sides (vacancies and curricula's).
- Using these statistics we propose a modification in a specific curricula. Our system will be able to make recommendations to remove, add or modify a specific subject. Based on literature review it is recommended to use Machine Learning techniques in order to create a recommendation system.

1.7. Thesis contribution

Our purpose is to create an automated model that makes comparisons between profiles and curricula. Actually there are manually researches made on comparisons between profiles and curricula, and that is considered the main contribution of our thesis.

Nowadays in our region there are accredited more than 500 study programs and still there is a gap between universities and industry. After this research universities will be able to use our recommendations in order to modify their curricula's.
Except universities, another contribution will be for the companies that request graduate students that have perception in: creativity, problem solving, communication, team work, etc. The advantage of our model is that it will be applicable in different countries with different companies. Even though our research will be made in a specific country: Macedonia, Kosovo or Albania, it will not be limited only for a country.

1.8. Methodology

The main purpose of this research is the suitability of university syllabus with the market demand in the region, as there is a major disadvantage in this regard. As we can see in the literature review, the adaptability of university programs to market demands is a problem that has begun to be addressed and is also required to work harder in this regard. Our main hypothesis H0 is "Automated Profile - Curriculum Compatibility Will Reduce Company Adaptation Time", and in order to prove this then an automated model should be created which will compile the level of suitability between the university syllabus and market demands. In order to arrive at such a pattern we will follow the steps below:

1) ***Profile – Curricula matching.***
2) ***Web Crawling***
    a. ***Curricula***
        i. ***Subject.***
        ii. ***Description.***
    b. ***Job Vacancies.***

3) ***Profile Clustering.***
4) ***Recommendation.***
5) ***Evaluation of the system.***

In the following we will describe each point in order to provide more detailed information on the methodology for how to implement such a model.

1.9. Thesis outline

Our research plan is divided in several phases which are accompanied with publications in conferences and journals.

Next we present the structure of our thesis documentation. As it is recommended from literature review and other PhD Dissertations, our thesis will contain 6 chapters: Introduction, Literature Review – State of the Art, Results, Automated model for comparison between profiles – curricula, Simulations and experiments of our model, Future work for other researches and Conclusion of our research.

In the first chapter, we are talking about the importance of adaptation between university curricula and labor market demands in the field of technology. Here are the latest researches that have been done in the labor market and forecasts for the future. This chapter introduces the problems that have motivated us to do this research and create an automated model that will be able to compare between the labor market and curriculum requirements offered by universities. Also in this chapter are presented the hypotheses and scientific questions of our topic that will finally get answers based on accurate results.

The second chapter contains the latest achievements in the fields to be applied in order to achieve automated model implementation. So it will be a state-of-the-art Data Mining field being divided into matching algorithms and web scraping and web crawling.

The third chapter contains the results that are derived from questionnaires that are disseminated in universities, private companies and government. With these results we will

be able to produce accurate results based on some views that have been in our interest in defining the topic. It is also important that with these results we will be able to answer some of the hypotheses and scientific questions that we have raised in the beginning.

The fourth chapter includes the implementation of the automated model, we will see how different profiles of study programs and labor market requirements will be created, later applying comparative algorithms that will make the current comparison of adapting labor market requirements to curricula offered by universities.

The fifth chapter contains simulations with experiments that will be made with the automated model. These experiments will be conducted in Kosovo and Macedonia in order to obtain a percentage of suitability in both countries.

Chapter Six is the conclusion and work that is recommended for the future. Given that this topic is of great importance, we will finally give the recommendations given to future researchers in order to continue work for the future and to achieve the perfection of such a model but also to meet the needs of labor market and universities not only in the field of technology but also in other fields.

2. State of the art

Nowadays the technology has reached a high point of development and research, and although WWW technology is found in its third decade, the application of Data Mining on websites remains one of the biggest challenges (Thakar, Mehta and Manisha, 2015). It is precisely the achievement and dynamics of the Web that has made the application of Data Mining in websites more difficult. Therefore, a technique of extracting information from websites needs to be applied in our case. According to (Sahu & Bhatt, 2017) Data-Mining offers powerful analysis techniques for different areas, especially appropriate techniques for the education field. Seeing the development of the education field, data growth for students, then Data Mining remains one of the most perfect techniques for achieving a successful research. When we talk about the field of education, then specific techniques are required for this field, as is the case with *"Educational Data Mining"* or *EDM*, where the main purpose is data processing in the field of education. Why an application of such a technique is important, it is shown by the increase of the data capacity in the education institutions' databases in our country and in the region, or otherwise known as Big Data. Just as Data - Mining works, EDM uses the same techniques such as: classification, clustering, and association rules.

a) The classification is used to classify the existing data based on the training set, and uses the template to classify a new type of data.

b) The clustering is used to divide into data sets that resemble each other, and those that differ from each other.

c) Association rules are used to detect links and dependencies between variables in large databases or Big Data.

In the third research question of the literature review we can see that Higher Education Institutions all over the world are facing difficulties in improving and managing their work and organization, this is also mentioned in (Sahu & Bhatt, 2017). Therefore, in order to achieve

this goal, Data Mining is considered as one of the most appropriate techniques which help these institutions make better decisions in order to improve their activity. Therefore, during the literature research, we have noticed several different analyzes that have been made by the Universities, such as: student performance, student challenges with the real world, analysis of offered literature, classification of graduates, and so on. But there is still no research on the appropriateness of the programs offered by Universities and labor market requirements, and such research is especially necessary in the countries of the region.

In the fourth research question of the literature review we can see that one of the biggest challenges of Data Mining is data extraction within dynamic web sites. It remains one of the biggest challenges, since in order to apply Data - Mining within a web site, we need to undertake more depth analysis and research of web content. There are many factors like website dynamics, increased security, increasing data from seconds to seconds, and so on, which have caused this to be the biggest challenge. The application of data mining on the web is commonly referred as Web mining. According to (Bharanipriya & Prasad, 2011) web mining is nothing else, but just a data extraction from the web site by applying Data Mining techniques. Knowing that the content of the websites is not only textual, but there are also other data such as documents, tabular, unstructured data, and semi-structured data, etc. When working with such data, one should work with unstructured, or at best with what it can be considered as semi-structured data. It is precisely these two forms of data that we can hope for in the case of our research. In order to retrieve data from the web, one should consider different apsects such as: *Source finding, Data selection, Generalization, Analysis and Visualization.*

The **source finding** aspect includes the phase of determining the source from which the information will be extracted, whether this *on - line* or *off - line* source.
Once the source has been found, **data selection** is required, in which we need to find the right tools and definitions in order to extract meaningful data.

The **generalization** is also known as the assessment phase. According to (Bharanipriya & Prasad, 2011) at this stage, machine learning or data-mining processes are applied in order to identify general patterns within certain web sites.

At the stage of **analysis**, data accuracy is measured based on certain data parameters.
The last stage is the presentation phase or the stage of **visualization**. At this stage, it is decided what form the presentation of the processed data will be presented.

There are different techniques and tools that can be applied for mining web content. Some of these tools are based on extracting data from the HTML structure of web pages, possibly by using CSS selectors to focus at particular kind of data. Other tools may at first extract the text from web pages, and then extract concepts from row text. These techniques are commonly referred to as *web scraping*.

Once the content we are interested is extracted, it needs to be processed in order to obtain meaningful results. In the following we will talk about classification as one kind of processing, where we will present the latest achievements in this field.

The fifth research question of literature review shows that web content has become the target of all researchers of the academy, while it is also mentioned in (Thelwall, 2001). In order to handle the content of the webpages we need a special application, which will go through several phases. Also, processing textual content of web pages may be costly in terms of processing time (Thelwall, 2001). There are nothing else but part of search engines in the form of an application that requires data in a certain, automated and regular way (Thelwall, 2001), (Adapure, Kale and Dharmik, 2014). Why crawlers are so important is the fact that search engines depend on them, enabling you to update information on up-to-date websites.

Unlike web crawlers, we can encounter the terms: robots, mice, worms, pedestrians, etc., and it is as old as the internet, since the first crawler was created in 1993. The shape of a crawler is moving from page to page using the website structure, and downloading its content to locally store it in a particular destination. Below we will describe the web browser architecture of how it works.

Web crawler is divided into three components: frontier, downloader page and repository. The first component used to keep the list of unrecognized websites so far. Initially, the content of the existing link is extracted from the web site and downloaded, then checked for

unrecognized links that are within the Web site. This link checking cycle continues until the frontier has been emptied. The second component or Page downloader is used to download web sites from the web that belong to links that are stored by the web crawler. As well as the third component or web repository is exactly where the content of downloaded websites is stored, but a large amount of data known as objects is stored here. The content that is stored is just HTML, all other types of data will be ignored. So these are the three components of how a web crawler works, but if we try to present the work of a web crawler through the steps then it looks like this:

1. *Splitting visited and unvisited links.*
2. *Placement at the frontier.*
3. *Selecting links from the frontier.*
4. *Extraction of the web page content that is linked to the boundary link.*
5. *Website analysis if there is any unrecognized link within it.*
6. *Placement of all unvisited links at the frontier.*
7. *Repeat step 2, until the frontier has been emptied.*

So this is the work of a web crawler through the steps of how it works, and if we want to describe it works like this: We have the program starting point, then checking the visited and unlisted links. Check if there is a termination of the application, if yes it terminates, if not it continues and selects a url link from the frontier, if there is no link it is interrupted. If the link exists then the content of the web page is extracted and the web site content is checked if there is any unvisited link, if another link exists it is placed in the parser, and after this step we have a rewrite of the cycle known as the crawling loop.

With the growth and development of the web, Internet users find it harder to track and receive all relevant web information without any automated process (Haddaway, 2015). Currently there are many applications that enable web scraping, and based on (Grasso, Furche and Schallhart, 2013) current applications that offer web scraping are offered as ad-hoc using complicated tools and languages. There is also a great variety in how these applications are provided, for example: some of them are owned by someone else, and some of them are more likely to be implemented. According to researchers in this area, there is a

great demand for such applications to be provided, since according to them, providers cannot expect to provide automated data extraction processes, since they are more concentrated in filtration and recommendations. In this paper we will present a project which is open source and is known as OXPath. This project is the continuation and expansion of XPath, with the difference that it contains four functions more than the XPath previous version. Through this module we can do web scraping in a very fast and efficient manner and with less cost, also the users of this module can experiment with the existing Firebug and FirePath tools (Furche, Gottlob, Grasso, Schallhart and Sellers, 2012).

On other hand we can see that the most important unsupervised learning problem is considered the Clustering. We can find many definitions about Clustering, but the most appropriate definition is "organization of objects into groups which has similar members in some way".

As a conclusion of literature review, we can see that all these tools are needed in order to create such a model. Also we can see that there is a need of creating an automated model for profile – curricula matching, which is our focus. So initially as a gap is identified: **Lack of automatized profile matching**. So figuring out the drawbacks that currently has matching profiles with curricula, where up to now we have been working manually in this direction and has cost over time, our model will be an automated model that will make automated profile matching. As sub - gap is identified: **Lack of automated design of reality gap**. So, as we know, currently Design Reality Gap is applied only manually, we will present it in an automated form to make a later comparison between the actual and the automated model.

2.1. Data mining

Latest achievements in Data Mining are huge, contributing directly to the development of different fields. Increment of data from time to time has made it necessary to apply Data Mining as a field in order to produce results as accurate and timely as possible. In (Han, Kamber, Pei, 2012) it is mentioned that we are now living in information time.

Lastly, we can easily see that we are dealing with digitalization of almost all areas, ranging from social to scientific ones, and that is precisely why we are dealing with such a large volume of data.

If we stop at social networks, nowadays we can see how from seconds to seconds we are dealing with increasing data in different formats: textual, photo, audio, video, etc.

On the other hand, if we stop at the scientific area, we know that the various exact sciences have reached in a high point of research by applying the technology and digitization their data. Therefore, this huge amount of data that is crossing the world day by day, it is necessary to apply professional techniques that enable us to process accurate results in a short time.

These data we are talking about, which grow dynamically at any moment may not be understandable to us when they are in the original format, but applying Data Mining techniques they are processed and converted into a format that is understandable to us. Therefore, as we can see, the main purpose of Data Mining is to adapt to the human language by becoming more and more inseparable part of us.

When we say inseparable part of ours, it should be noted that thousands of applications today function as an inseparable part of the human world by giving different people recommendations on different activities. These recommendations given to different people through various applications today are made possible through Data Mining techniques. And it is precisely the collection and processing of various information that has influenced these applications to give us recommendations and conclusions through the Data Mining techniques.

A concrete example is the case of medical analyzes, where digitized devices are able to produce results on our health condition based on earlier samples of different persons that are stored in the system. The greater the capacity of these data stored in the system, the greater the precision of the result determined by the digital system. This is an exact example of how Data Mining converts data into a format that is understandable to the human world as well.

In (Han, Kamber, Pei, 2012) it is mentioned that evolution and technology development has occurred precisely because of the tremendous development that Data Mining has suffered. This dependence between technology development and Data Mining lies in the fact that the

data available today on the internet is of a great variety. The data that Data Mining can handle are: textual, numeric, structured data and unstructured data, graphical data, and data that are distributed through web systems.

In the following we explain how Data Mining has found the application on each of them.

2.1.1. Textual Data

When we talk about textual data, the Data Mining application process in the text is known as Text mining or Text Data mining. Text mining includes the process of text processing to the presentation in a professional and high quality. The process of text processing through Data mining techniques is the process of removing a word in the text, adding a new word, or completely structuring a document based on a model that the system possesses as a training model. And once the text has been processed, then it becomes ready for presentation in a format that is understandable to the human eye.

So the main purpose of the Data mining application in text is to convert textual data into data that are readily available for analysis. A simpler method for Text Data mining is converting documents from hard copy to electronic format in order to enable the application of Data mining techniques. Some of the areas that directly depend on Text mining are state intelligence, search engines, publishing houses, social networks, and so on.

2.1.2. Numeric Data

Creating different models that are able to provide predictive results based on preliminary results is precisely the application of Data mining to numerical data. Nowadays, many predictive analyzes are needed that help us to make decisions based on the result that the system exits. A method by which we can make predictions based on preliminary results is through Bayesian. Through this method we can make data sorting by dividing the data into two or more groups that the system classifies. Even in this case, the more data that the system possesses the more accurate the classification that the system determines. Some of the domains where Numerical Data mining has found application are: math's, medicine, physics, chemistry, and so on.

2.1.3. Structured and unstructured Data

The data that can be found in a database can be structured and unstructured. Structured data are the data that are organized in the best possible way. While unstructured data are data that do not have an organization and a structure. When we are in structured data, it's easier to apply Data mining techniques as we have tables that are related to one another, so when a given data is needed then we know exactly which table and we which line should we ask for it. While the main problem and concern lies in unstructured data as we have no information on where to find it. However, through Data mining techniques we can create intelligent systems using unstructured data and creating in this form of structured and understandable human data forms.

2.1.4. Graphical Data

A kind of data which is difficult to process are the graphical data due to the complexity they possess and because of the difficulty of conversion in data for analysis. Currently in almost all areas, the results are converted into graphs in order to be more understandable during different presentations. With Data mining techniques we can also process these data in order to obtain desired results. Data mining application in graphical data is known as Graph mining, and unlike other Data mining techniques this technique is less accurate (Romanowski, 2009).

2.1.5. Web Data

Nowadays companies of all areas promote themselves through the web platform. Apart from the information that should contain a website, designers today are also focused on the dynamics that the website incorporates, including information as part of design and multimedia. It is precisely this fact that has troubled the work of extracting and processing information from web sites. Our automated model depends exactly on the websites, as the main information will be extracted from them. And the best solution is to apply Data mining techniques to the web. Two techniques that enable us to extract this content are Web Scraping and Web Crawling, which we will later discuss in more detail.

2.2. Data mining evolution

In the following we will present the evolution of Data mining and its history of how it has reached this peak point where it is today.

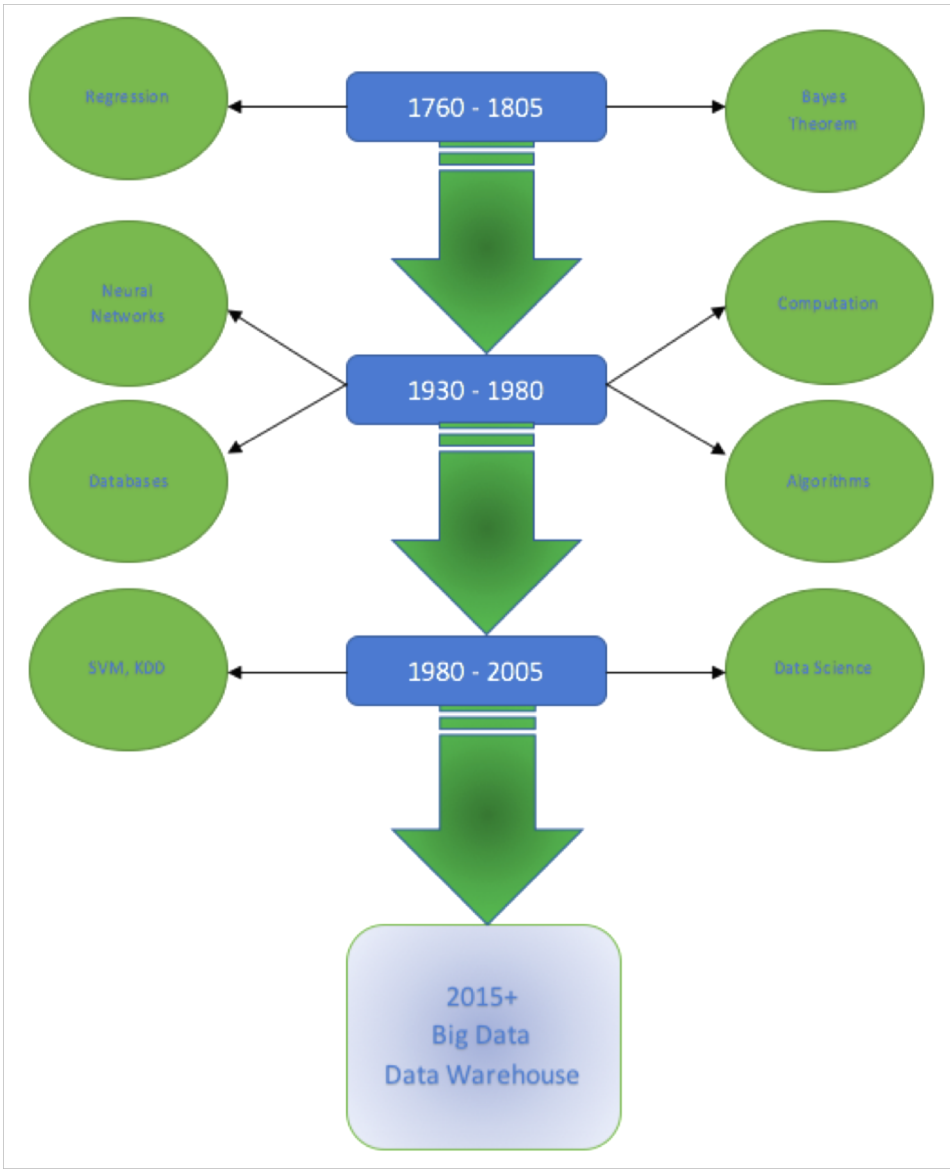


Figure 2. Data Mining evolution

In Figure 2 we can see how Data mining has been developed since the 17th century, where from 1760 to 1805 we have the discovery of two theorems: Bayesian and Regression. Currently, through these two theorems, science has reached high discovery by yielding accurate results and predictions regarding certain cases. We can also note that in the 19th century there have been great discoveries and great advancements in the Data mining field.

Specifically, from 1930 to 1980 is the period when the first databases were created, also the Neural Networks was discovered. It is precisely this time when important algorithms that day-to-day use are being discovered, and since then, they have been sophisticated and advanced in different languages. Also this period is well known for the digitization of these discoveries since all the theorems that were discovered and all the algorithms that were discovered began to be applied through the computer, directly affecting them at the time of their implementation and in the accuracy of the processed results.

Not by chance that these algorithms were applied to the computer, since at this time the volume of data that was stored on the Internet began to increase as the Internet has begun to evolve. Also, in this period, from 1980 to 2005, other discoveries of various algorithms that are known today and are very applicable were made, such as SVM, Deep Learning, KDD, etc., Therefore, this period is also known as the period of the Data Science, at that time scientific methods began to be implemented to produce data and to present the data in a human-accessible format, whether structured or unstructured.

In Figure 2 we can also note that the last period since 2015 is known as the Big Data period, where the data format that is stored is petabyte above. So this is exactly the time when different businesses and corporations began to use technology in order to create intelligent systems that represent their business in the best possible way. Also in this period, the Data Mining application is started on giant and data sets known as Big Data. Also, the Data Mining application at Big Data has made it possible to make data processing in a precise and fast way. Finally the areas where Data mining has found application on Big Data are different, such as: Medicine, Education, Social Science, Marketing, Finance, etc.

2.3. Clustering

The division of data into certain groups based on the similarity of objects is known as Clustering. Like other Data mining techniques, clustering is one of the techniques that has managed to develop alongside other techniques. Developing and advancing clustering techniques has also led to the development of many other areas such as medicine, education, finance, marketing, machine learning, etc.

The way clustering works is by creating as many groups of objects which are the same with each other. The greater the likeness of being within a group, and the bigger the difference between the groups, the greater the clustering accuracy.

In addition to clustering there are many other techniques that make data collection in different groups, but in some literature we can also find that clustering can also be known as a form of classification. We say that it is known as a form of classification, as it divides into data groups by classifying them into different groups. But distinction from the classification technique that classifies the data into different groups, and when it comes to a data that is classified as a new group, clustering these new data tries to group them with the actual data being compared to the most similar figure. In the following we will show graphically how clustering works.

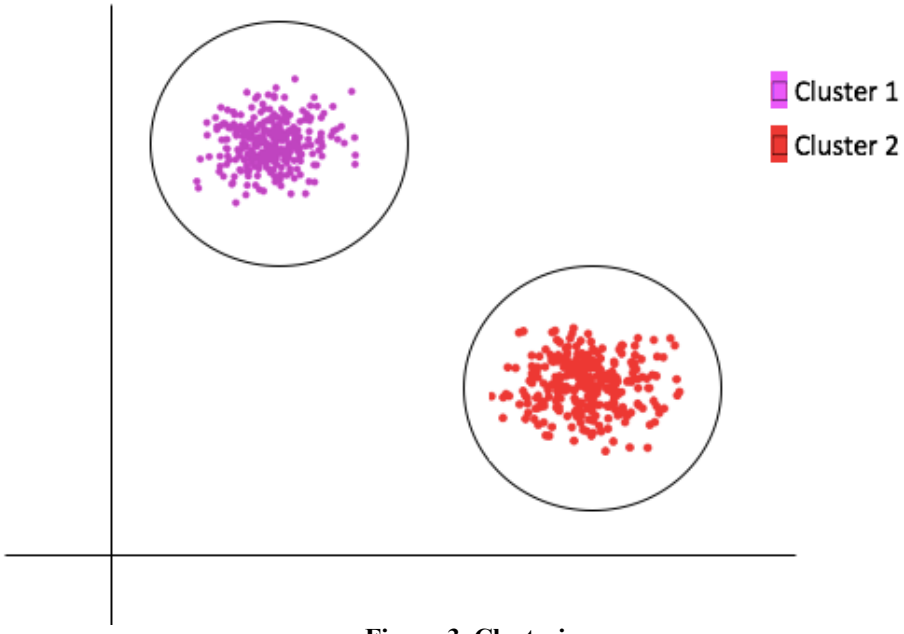


Figure 3. Clustering

In Figure 3 we can see how the clustering technique is divided into two sets of data that are located in our dataset. These data are divided into two groups that differ with each other, but with data that are close to each other. In our case the data are divided into two groups only because of the difference between them, but in other cases we have more groups. Also in Figure 3 we can see that we have a high level clustering because the data are completely separated, but there may also be times when the data cannot be completely separated but only join to the group which is more closest. When we are in clustering, we can point out that clustering types are: well separated, center - based, contiguity - based, density - based and conceptual clusters.

2.3.1. Well separated clustering

The type of well-separated clustering is the type when the data contained within a dataset are completely separated into the groups that are identified. And when the data is fully approximated with the data that are within a group then we can conclude that we deal with well separated clustering. Similarities between data that are within a group, and the distinction between two or more groups at clustering is measured by distance. Therefore, when the distance between the two groups is greater than the distance between data that are within a group then we can conclude that we are dealing with well separated clustering. The best example of well-separated clustering is presented in Figure 3.

2.3.2. Center – based clustering

In cases where the data division is not the maximum but can be considered at a satisfactory level, then we can conclude that we are dealing with center - based clustering. This kind of collection is worked on the average, so an average of the data is collected within a group. And the distance that is compared is the distance between the data and the center or average that is defined in this cluster. In other words, all data is approximate to the average of the group, and if the distance is smaller than the average, then the data belongs to the given cluster.

2.3.3. Contiguity based clustering

In this type of clustering, an important role has also the separated files that are within clusters created by the system. We cannot consider that there is any close approximation to the average or all the data that are within a cluster, since in this case it is sufficient for the given data to have a similarity as small as at least one of the other data not that group. In other words, if a given match is more consistent with any data in a group than the data in the other group, then the given data is classified and entered within that group.

2.3.4. Density based clustering

Density based clustering is an algorithm that was proposed by Ester, Sander, Kriegel, and Xu in 1996. This algorithm works by dividing the data into different clusters, but within a cluster by merging the data that have great similarity and those that have the low similarity.

The division takes place in regions known as high-density regions and low-density regions. In other words, although the data is within a group, if they have a distance with the data within the group then the system classifies them and places them on low - density regions. But still the distance between the data within the two regions should be smaller than the distance to the other cluster created by the system. It is also worth noting that Density based clustering algorithm is one of the most useful and widespread algorithms in the Data mining field.

2.3.5. Conceptual clusters

In previous types, the data contained within a cluster had no relation to the data that were within another cluster. In Conceptual clustering we have a different situation, since the data of one group contain some of the features of the other group. In other words, if we have two clusters that are divided into groups, and the data that has the properties of both groups then those data are placed in the middle. Also, if we are dealing with other data entering the system, then they will be placed in that region where it shows that the given feature has the features of both groups that are created in the system. As far as group creation is concerned, there is no limit to how many groups can be created.

2.4. Matching algorithm

Comparison and finding the suitability between the two elements is required in virtually all fields (Melnik, Molina, Rahm, 2002). In order to have clearer how the algorithm functions, we will give a brief explanation of how matching algorithm will be applied to our model.
As mentioned above, our model is an automated model that will make a comparison between labor market demands and curricula offered by universities in the field of technology. So far we have mentioned all the techniques and methods that will be applied in order to implement our model, and we have reached the matching technique. So knowing that our model will create different profiles of labor market demands and university curricula, then it is imperative that these profiles be compared and find the level of adjustment of between them. This level of adjustment will be found by applying the matching algorithm to these profiles that are created using clustering techniques for which we have provided clarifications in the previous chapter. In the following we will graphically illustrate how the matching algorithm works which will be applied to our model.

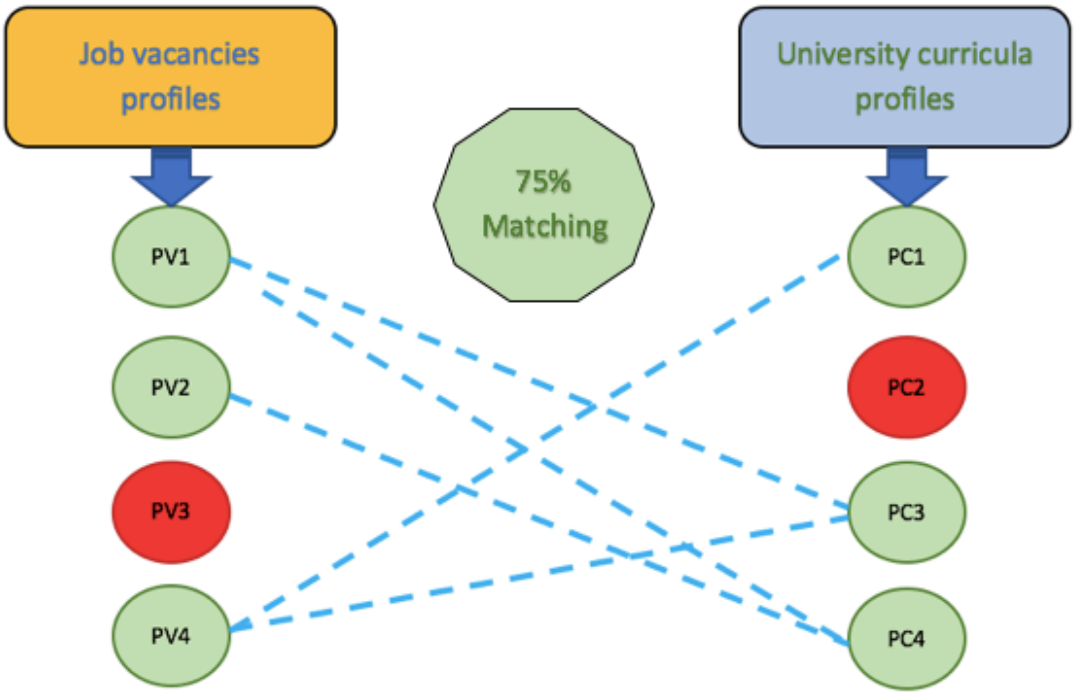


Figure 4. Graph Matching Algorithm

In Figure 4 we can see how the matching algorithm works, which in this case is illustrated how it will be applied to our model. As we can see on the left hand we have job vacancies profiles that are named from PV1 to PV4, and under the right hand we have university curricula profiles that are named from PC1 to PC4. After applying the algorithm we can note that the PV1 has PC3 and PC4 adaptability, as well as PV2 has PC4 compatibility, and PV4 has PC1 and PC4 compatibility. As we can see profiles vacancies 3 and profile curricula 2 have no relation to any of the profiles. And what results from this result is that profile 3 of the job remains uncovered, and in that case our system will be able to give recommendations that this profile should be covered so that a new curriculum can be added.

Or in the other case we have profile curricula 2 which also does not have any links to any of the profiles of work, then we can conclude that this curriculum is not needed in the labor market, and our system will be able to make recommendations on the changes that need to be made in this curriculum so that it responds to the demands of the labor market.

Also based on the results that derived our algorithm after the application, it can be concluded that out of 4 profiles, 3 of them respond to labor market demands or 75% is the level of adjustment between the demands of the labor market and the curricula offered from universities in the field of technology.

2.5. Web crawling

The program that allows us to automate browsing through websites that are active is known as web crawling. The way the web crawler works is by checking webpages that are active in certain phrases that we define by themselves. Some websites today use web crawling as a perfect way to keep their web pages updated.

The other way web crawlers work is by visiting the web pages automatically and downloading them to a local disk that we set as a destination to maintain the content. The content of the webpage we can download, starting from static to dynamic, but depending on the dynamics of the webpage depends also on the script we need to build in order to have as much information as possible. Below we will graphically show how the web crawler works.

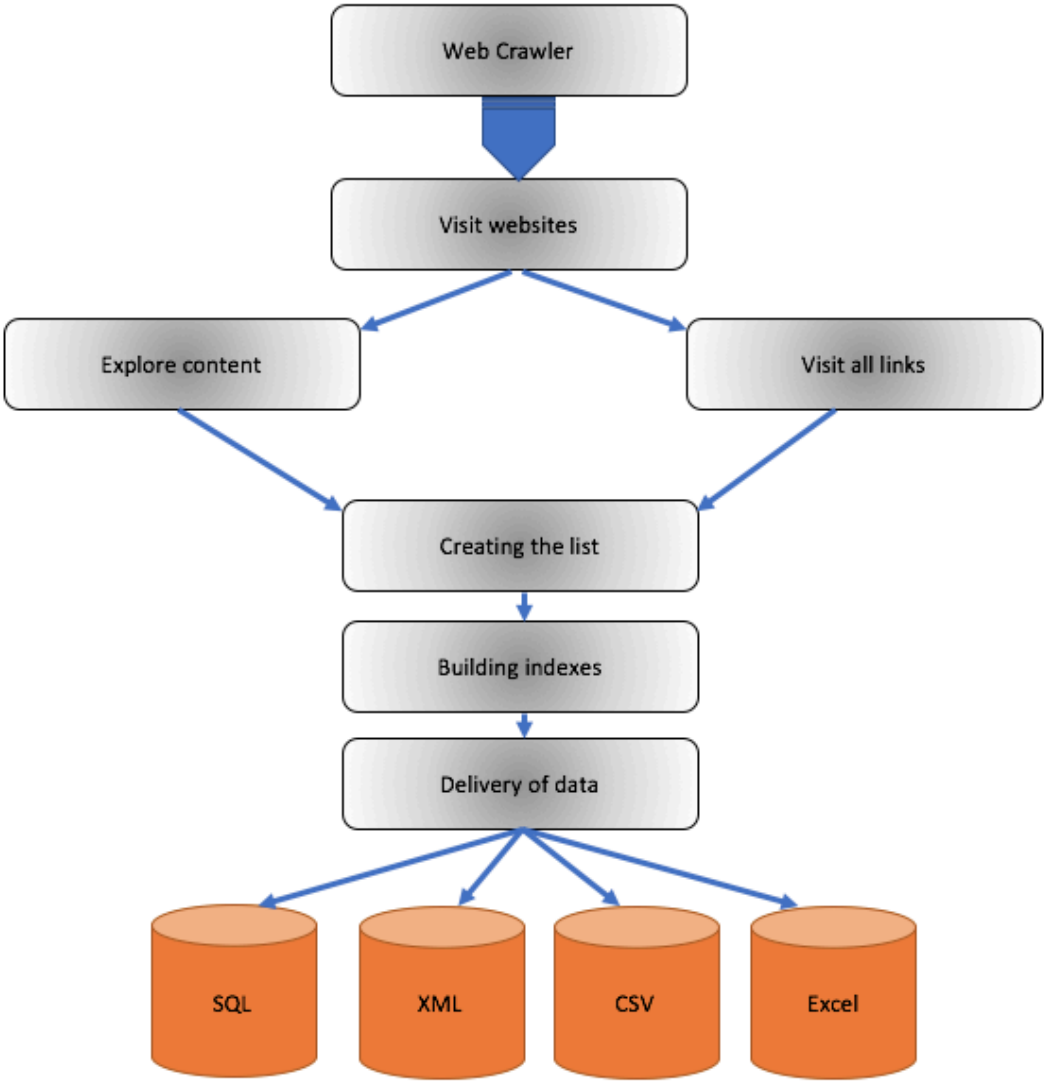


Figure 5. Web crawling

In Figure 5 we can see that the first step in which the web crawler application passes is the site visit that we define at the beginning. After a visit to the web, the crawler makes an exploration of the content based on the phrases that we have assigned at the beginning, and at the same time the crawler visits the other links that are defined within the webpage.

Once the textual content has been extracted, the crawler creates lists of those web pages, in order to create indexes that will be downloaded later.

Once the indexes are created then the last step is to save the data, or as we have presented it as delivery of the data, where the data we extracted from the content of the webpage can be stored in any format that we have need. The most used formats that can be converted to the data extracted from the web site are: SQL, XML, CSV, Excel, etc.

Based on this, the main reason for web crawling will be applied in our research is to extract information that is published by some websites on vacancies. These data will be extracted in certain keywords, where their clustering will also be based on them. Initially in the region there are web pages that contain data on university programs, these data also contain descriptions of the subject and description of the study program.

Based on this information, will be the clustering of study programs that are offered by universities. These later clustering will be used to match the market requirements.

On the other hand, Crawling will also be applied on web sites that provide information on competitions offered by different companies. Also, the web crawling application will be made to specific keywords, knowing that each contestant has additional information on the specific position requests.

So, every position that is required to keep information in addition to the required position, the applicant must have knowledge in several different areas such as: programming, databases, networking, etc. All of these descriptions of later positions will be used to gather different positions.

3. Higher Education and private companies in Kosovo

We have noted the importance of higher education in all publications that are related to this field. Rather than affecting the greatest employer's ability, some studies see higher education as one of the leading factors that directly affects the style of life that we do.

Therefore, today's demands have increased significantly for skilled people, and prepared in complex areas.

It is precisely technology that has affected the lives of people today to be more problematic and more complex, making the vast majority of information available today in the electronic format. And precisely this large volume of information, and the exceeded number of technological devices, has made demand for the labor market in the field of technology to be increased every day and more.

In Kosovo, this problematic is even greater by counting some factors that have affected, such as the economic situation, high unemployment, non-compliance with labor market demands, etc.

In order to accurately identify the problems and deficiencies that have high education in Kosovo, we tried to get answers to various issues from public and private universities in Kosovo.

Though some of our hypotheses and scientific questions have come to be confirmed through these analyzes. We have also managed to identify the real needs that the market possesses in relation to universities. The results we have received are divided into three sections: academic staff, students and university quality.

As for the academic staff, we have first tried to identify the academic degrees and degrees that academic staff possess within public and private universities. Also in this section we have analyzed the opinions of different professors on the curricula offered by these universities, the professional preparation they offer and the identification of the deficiencies that the students possess during the studies.

In the part of students we have also analyzed their professional preparation which they can take during the studies, the difficulties they have during their studies and their participation in mobility.

While in the quality part we have analyzed the career centers that the universities possess, the participation of students in different projects such as Erasmus, the possibility of involving businesses and students in the curriculum working group, etc.

Taking the results is done in electronic form by designing forms of questions of different nature. Then, the results we have received have been processed with the SPSS application. Through this application we have made various analyzes, using cross-tabulation techniques in order to find the interrelated results and variance among the different variables.

In the following, we present these results in graphical form, giving an explanation for each of them in detail. Initially we start with general results, to go on with detailed analysis of higher education in Kosovo.

3.1. Academic Staff

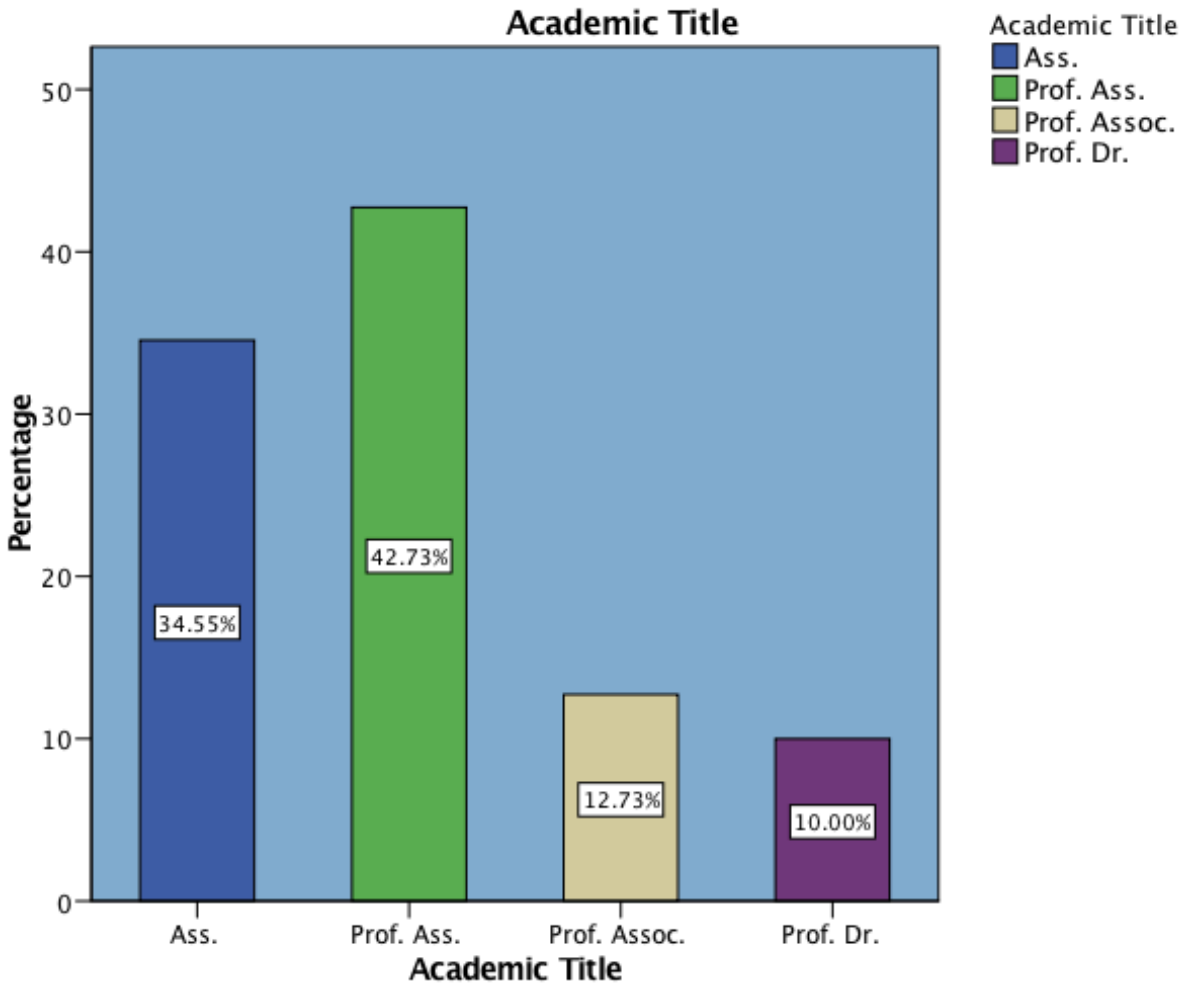


Figure 6. Academic staff

In figure 6 we can see the academic staff located in public universities and private colleges in Kosovo. As we can see , over 80% of the academic staff are assistant and assistant professor, while the other 20% are associate professors and full professors. Based on the recommendations that the higher education institutions receive from the Kosovo Accreditation Agency, more and more professors are required with the last two in ordinary and ordinary academic grades. Therefore, based on these data it can be considered that the situation is worrying as regards the accreditation of new education programs in the field of technology.

3.2. Preparation of students

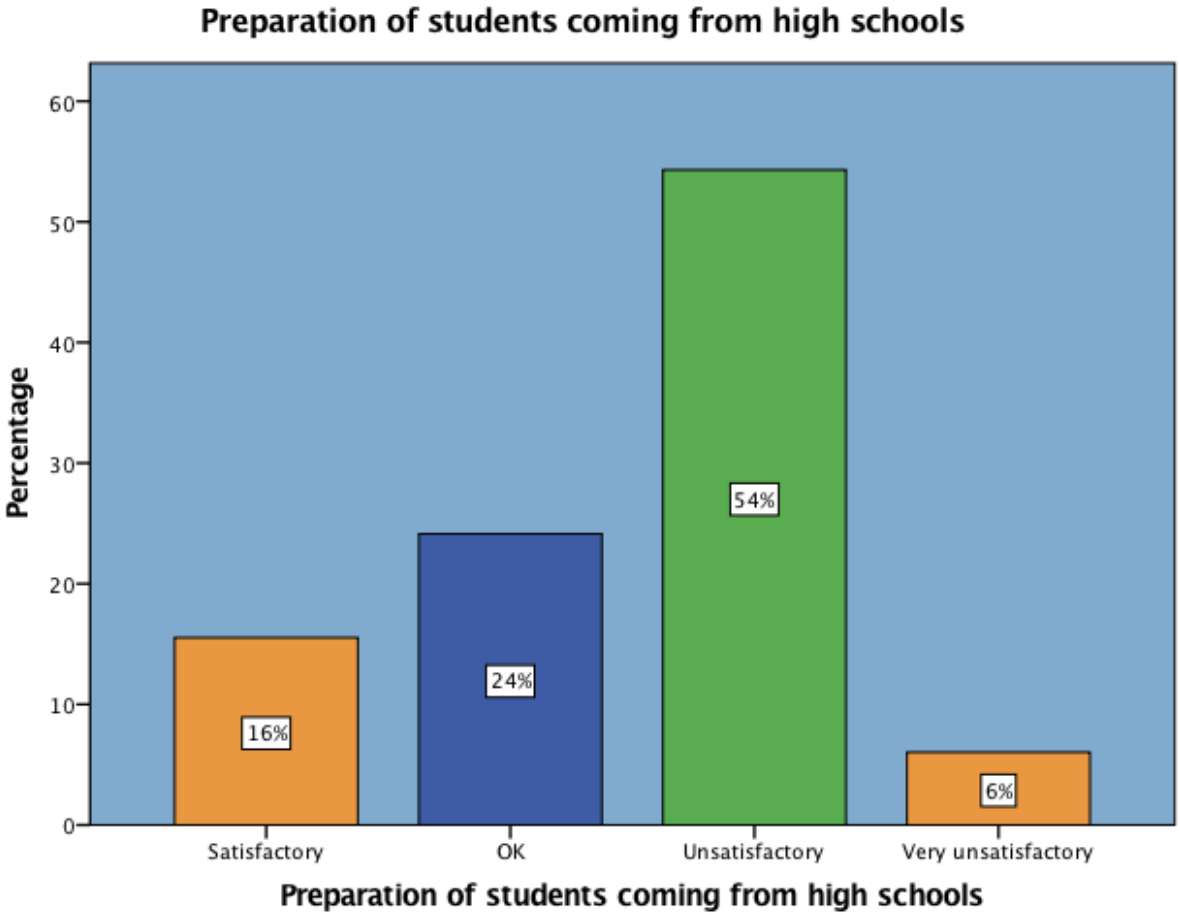


Figure 7. Preparation of students coming from high schools

In Figure 7 we can see what is the preparation of students coming from secondary schools. The responses are divided into five groups: Very satisfactory, satisfactory, OK, unsatisfactory and very unsatisfactory. Based on the results obtained from public and private universities,

we can see that 60% of responses have to do with an unsatisfactory level, and 40% is of a satisfactory and average level. No university has stated that the level of student preparation is very satisfactory. This is also a factor affecting the preparation of students for the labor market, as the preparation of students also depend on the knowledge they have gained during the secondary and lower secondary studies.

3.2.1. Preparation of graduated students

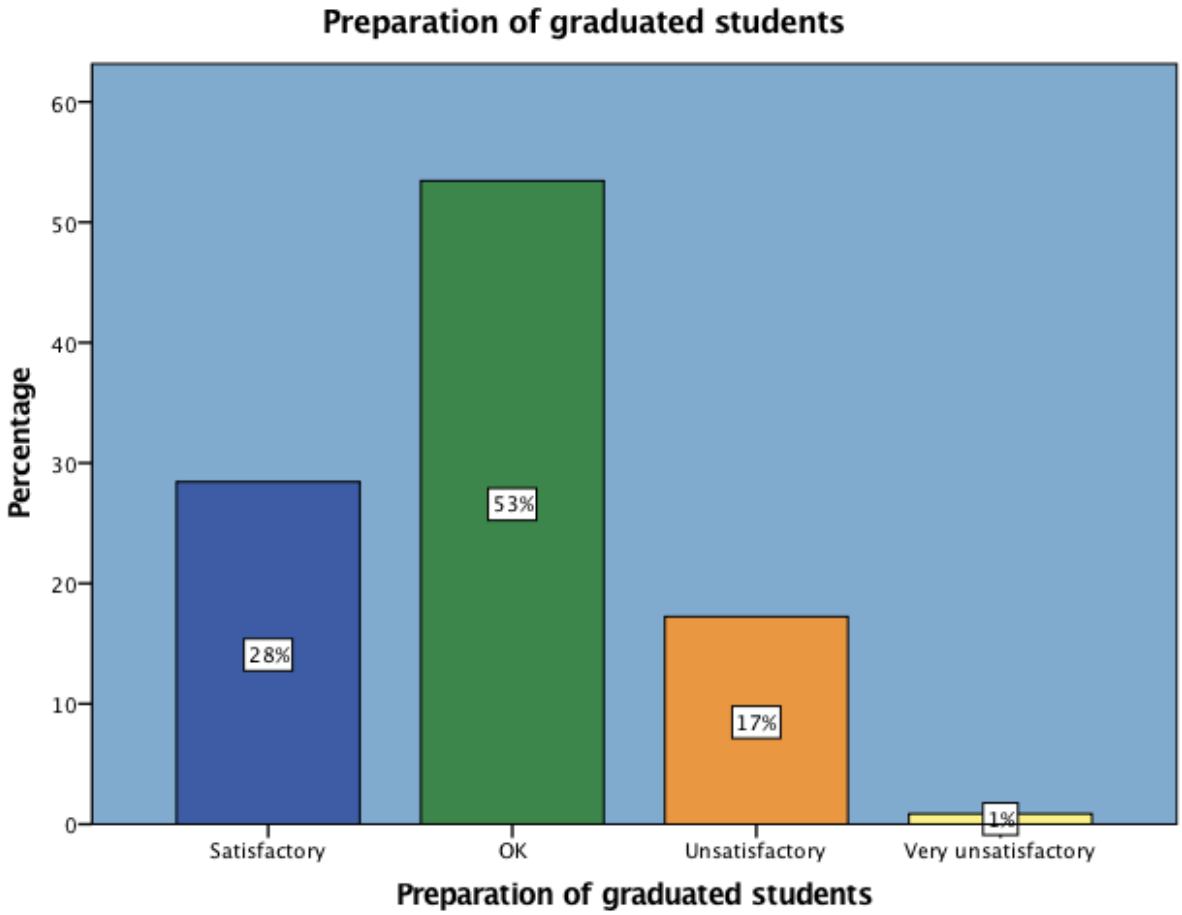


Figure 8. Preparation of graduated students

In figure 8 we can see that the level of student preparation is not the same as that of the students coming from and with the university. So we can see that 80% of respondents have given a satisfactory level of student preparation after completing studies, while 20% of them have given an unsatisfactory level of student preparation after graduation. Although the difference exists, the level of discontent is still high, and to see if there is any dependence on what the students are getting with what they graduate, we will present the analysis done between these two variables.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	58.319[a]	9	2.823E-9
Likelihood Ratio	53.807	9	.000
N of Valid Cases	116		

a. 9 cells (56.3%) have expected count less than 5. The minimum expected count is .06.

Figure 9. Students preparation variance

In Figure 9 we can see that from the analysis done between the preparation of students coming from high schools, and the preparation with which they get to graduate from university have dependence on each other.

The variance is very large since significance is compared to **α = 0.005**. In our case, the significance is **2.823E-9**, which is much less than 0.005, so we conclude that the preparation that students can take during their studies and with which they graduate depend directly on the preparation with which they come from high schools.

3.3. Level of students who continue their master studies

The number of students in second cycle studies is of particular importance to the university, as there are several factors influencing this. First, if the students who complete the first cycle studies and continue the second cycle studies at the same university, it results with great pleasure of the students with the curricula that the university offers. When we are in the curricula and their adaptation, it is not required that only in the first cycle studies have the adaptation of the curricula with the requirements of the labor market, but in the three cycles it is required to have an adjustment. Below we present the level of students completing the first cycle studies and continuing the second cycle studies at the same university. The results surveyed by universities are also based on the statistical data that universities have for their alumni.

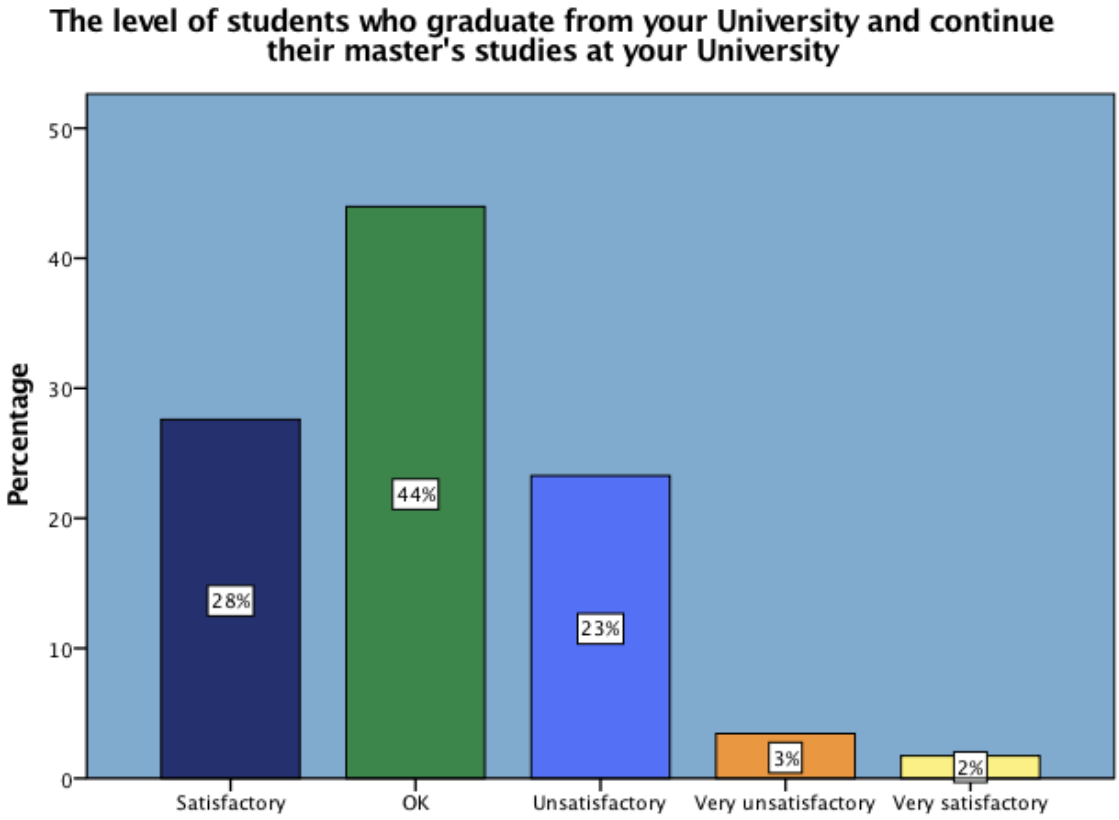


Figure 10. Master students level

As we can see in Figure 10, the level of students who graduate in university and continue their second cycle studies at the same university is over 75%, while 25% of them have declared that this level is unsatisfactory, it turns out that students finish master studies at other institutions. On the other hand, regarding the continuation or non-continuation of second cycle studies at the same university is also the adaptation of university curricula to the labor market requirements.

In the following we graphically illustrate the employment of students after their graduation. Also, these statistics regarding student employment are based on realistic data that universities have through alumni associations.

3.4. Employment of graduated students

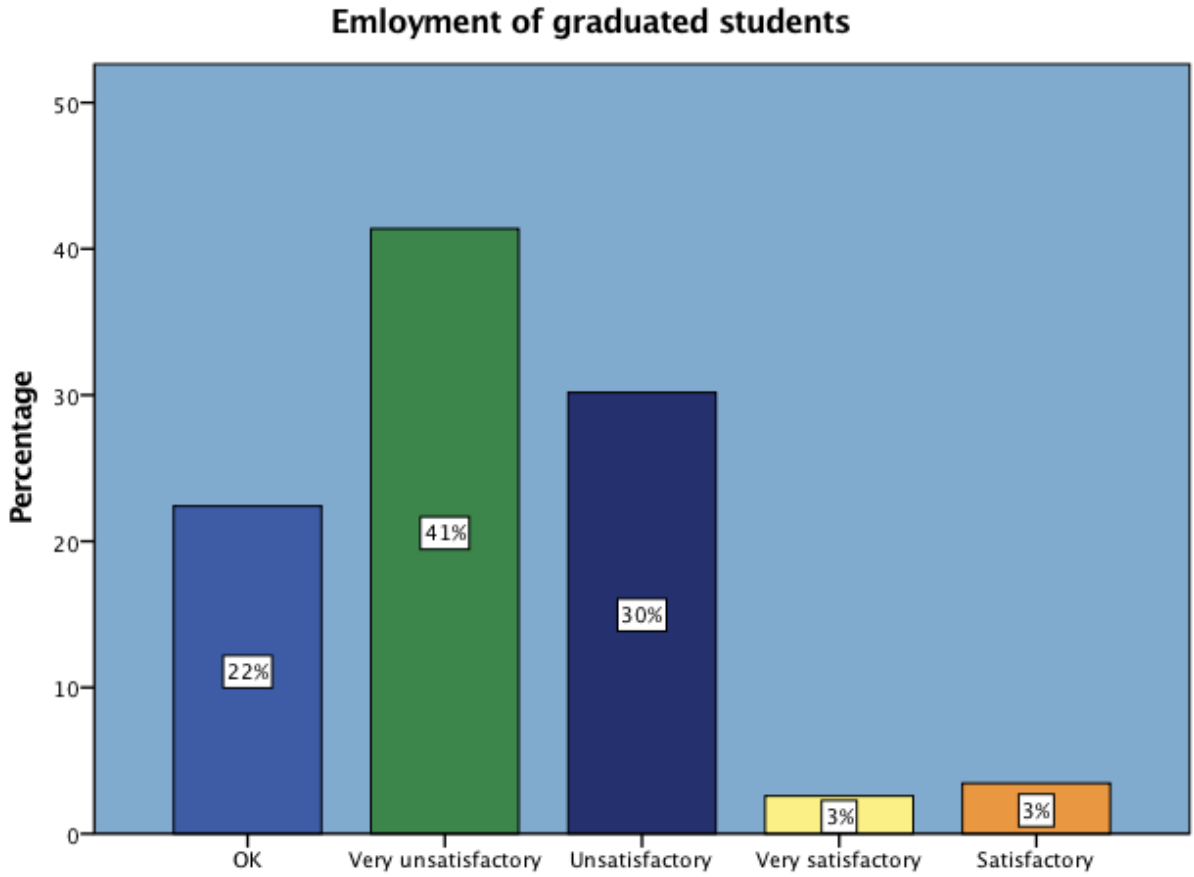
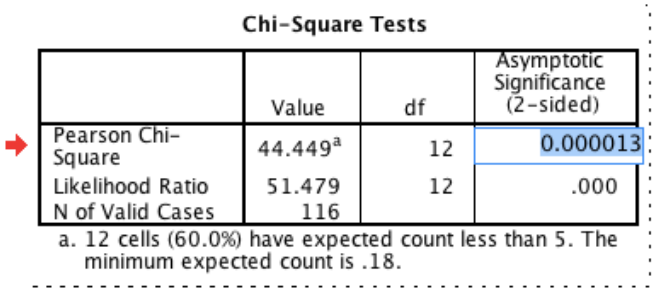


Figure 11. Employment of graduated students

In Figure 11 we can see that the level of employment of graduate students is at an unsatisfactory level.

Based on the data extracted from private and public universities, 70% of them are at an unsatisfactory level with regard to employing their students. While the rest of 30% stated that the employment level is satisfactory. A high percentage of unemployment of graduate students results precisely in the non-matching of labor market demands with the curricula offered by universities. In the following, we graphically present the analysis made between student employment and the university agreement with private companies. Part of these agreements is also the participation of representatives of private companies in curriculum development groups offered by universities. Besides the representatives of private companies in this drafting group are also part of the student, and precisely the main purpose of business and student representatives is to give proposals as to what should be included in the curricula offered by universities.

3.4.1. Employment of graduated students and cooperation with private companies



Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	44.449[a]	12	0.000013
Likelihood Ratio	51.479	12	.000
N of Valid Cases	116		

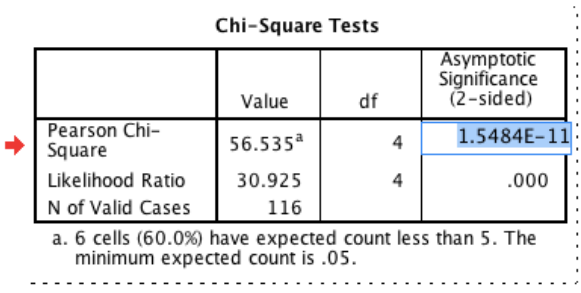
a. 12 cells (60.0%) have expected count less than 5. The minimum expected count is .18.

Figure 12. Employment vs University cooperation with private companies

In Figure 12 we can see that there is a strong link between student employment after graduation and the agreements that universities own with private companies. As can be noted, the significance is **0.000013**, which can be considered to be on a reliable link level between these two variables.

But are there different data in Kosovo and Macedonia, and is there any dependency between student employment and the state that they will come from in the graph below. The comparative data we have used come from the state of Macedonia as our questionnaires are also disseminated in this country so that we can make comparisons between important variables such as the employment of students after their graduation.

3.4.2. Employment of graduated students in Kosovo and Macedonia



Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	56.535[a]	4	1.5484E-11
Likelihood Ratio	30.925	4	.000
N of Valid Cases	116		

a. 6 cells (60.0%) have expected count less than 5. The minimum expected count is .05.

Figure 13. Employment of graduated students vs State

In Figure 13 we can see that we have a strong dependence on student employment after graduation in Kosovo and Macedonia, whereby students in Macedonia get employment after graduation at a better level than in Kosovo.

3.5. Areas where students face difficulties

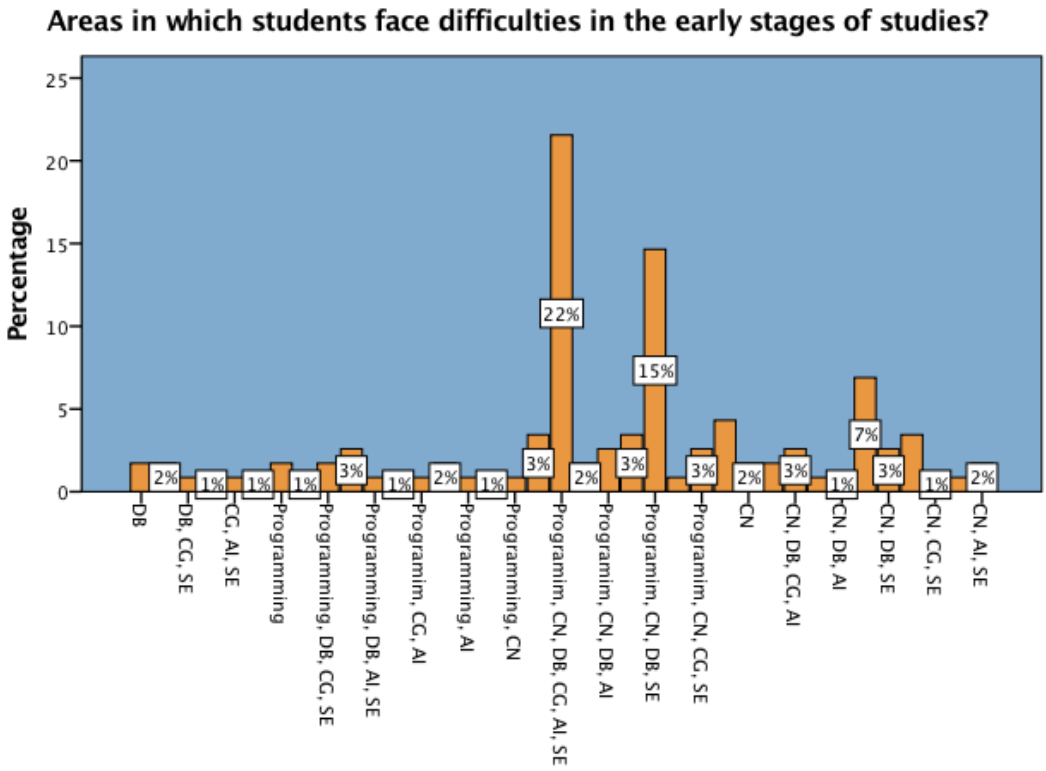


Figure 14. Difficulties that students face in the early stages of studies

As can be seen in Figure 14, the inclusion of vocational subjects that are the basis for studies in the field of technology has been included. These subjects are listed in this form: Programming, Computer Network (CN), Databases (DB), Computer Graphics (CG), Artificial Intelligence (AI) and Software Engineering (SE). Based on the results obtained from public and private universities in Kosovo, there is an exchange between all the subjects. However, during the responses the academic staff and students have provided, it has been possible to combine all subjects into different groups in order to determine which class of subjects are faced by students in the early stages of study. The group that has most of the answers received from academic staff and students is the group of subjects: **Programming, Computer Networks, Databases, Computer Graphics, Artificial Intelligence and Software Engineering**. So there are all subjects that have been in combinatorial choices that are selected as subjects where students are experiencing difficulties. These difficulties occur precisely for the fact that the students are coming from secondary schools, which unfortunately, were unhelpful in the universities. The other reason students encounter in the adversity is also due to their professional preparation offered by the university.

3.6. Professional preparation of students

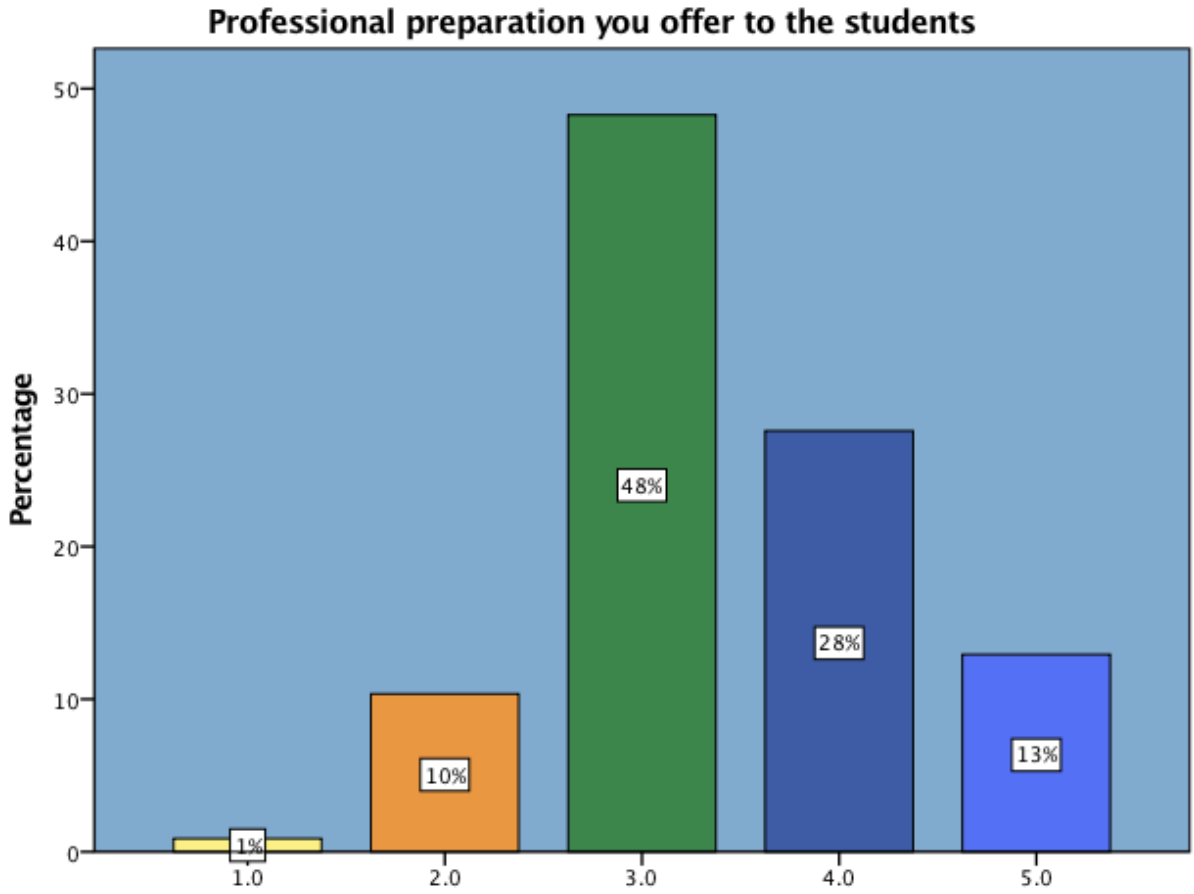


Figure 15. Professional preparation that is offered to the students

In Figure 15 we can see the professional preparation that is offered to students during the first and second cycle studies. The range of responses to how the academic staff responded to the professional preparation offered to students was from 1 to 5, with 1 resulting in very low preparation and 5 being too high.

From the results that have been processed, we see that the professional preparation offered to students is above the acceptable average, as 50% of respondents said that the professional preparation offered to students is at the third level, and 40% fourth and fifth level.

With only 10% responded that vocational preparation is at an unsatisfactory level resulting in curriculum improvements, academic staff advancement, and practices that students have to end up in private companies as a result of agreements the universities possess. How much the professional preparation for student employment affects, we will graphically present the analysis that is made based on these two variables.

3.7. Professional preparation vs employment

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	45.340[a]	16	0.000123
Likelihood Ratio	41.508	16	.000
Linear-by-Linear Association	7.558	1	.006
N of Valid Cases	116		

a. 18 cells (72.0%) have expected count less than 5. The minimum expected count is .03.

Figure 16. Professional preparation of students and their employment

In Figure 16 we can see that the professional preparation of the students is directly related to the expectation that they have to be employed. Based on the analysis of the results the value of significance is **0.000123** which is a reliable value if compared to the limit **α = 0.005**. Therefore, we can conclude that the professional preparation of students is related to what the students expect to get employed to graduate during graduation.

Another factor that plays a role in preparing students for the job market is the career center, and we will graphically illustrate how much the universities in Kosovo possess such a one.

3.8. Career center at Universities

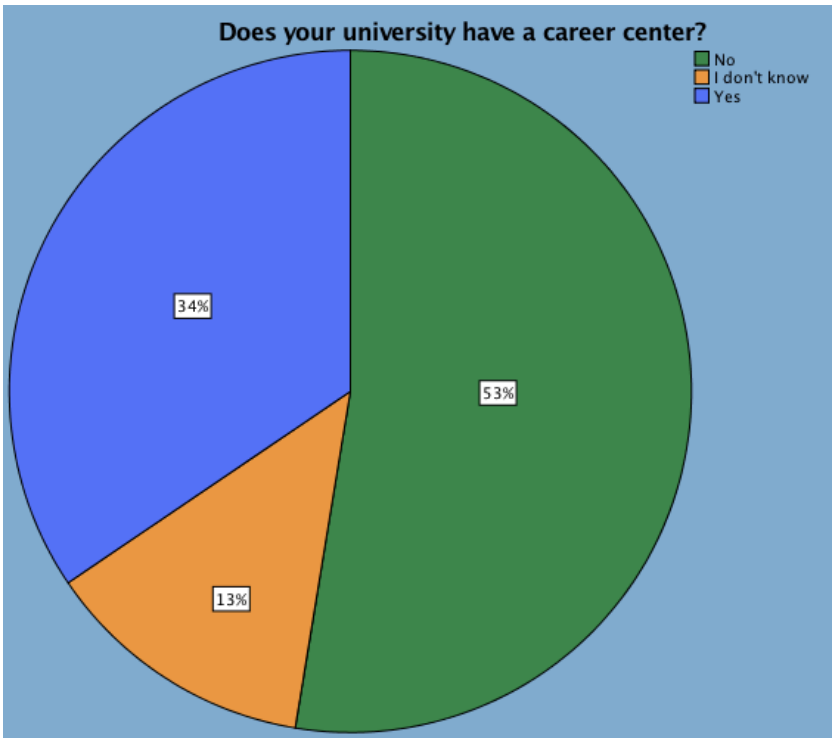


Figure 17. Universities Career Center

The importance of the career center is shown by the fact that all the employment opportunities offered by private companies reach the students through the career center. How much do they possess such a center in universities in Kosovo can be seen in Figure 16. According to this chart, which was built on the basis of the answers received from the academic staff of the universities, 53% responded that they do not possess such one, 34 % responded that they own such one, and 13% do not know whether they own or not. Based on these statistics, universities in Kosovo pay little attention to such a fact, negatively affecting the promotion of their students. In this form universities do not even possess realistic data on labor market demands and on the opportunities that happen to their students. In the following, we graphically illustrate how universities are aware of the labor market demand that exists in the field of technology.

3.9. Market demands in Kosovo

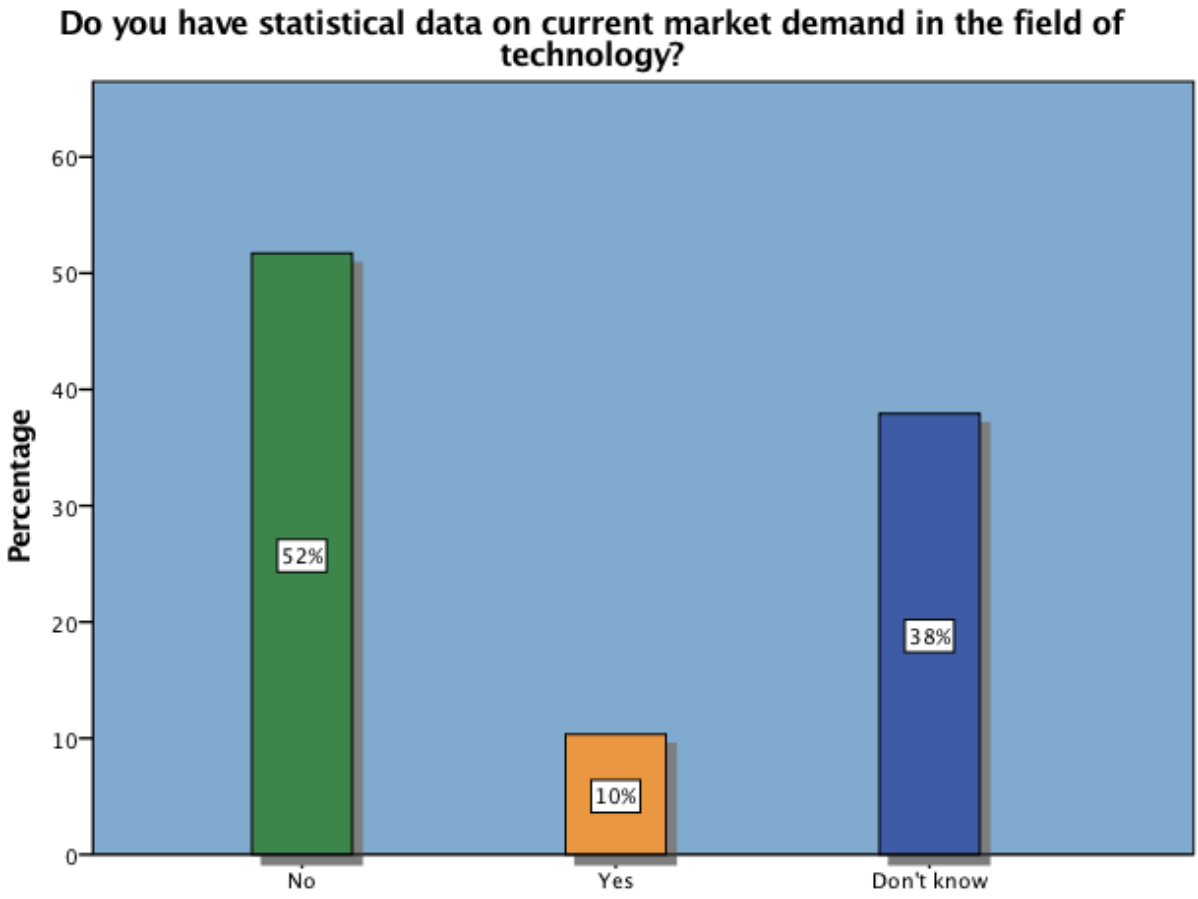


Figure 18. Market demands in Kosovo in the field of technology

Based on the responses received from public and private universities in Kosovo whether they have knowledge of the labor market requirements or not, 52% of respondents have replied that there are no statistical data on the labor market demand in the field of technology. Also 38% responded that they did not know if they had information or not, while a very small percentage of 10% responded to having information on the labor market demand in the field of technology.

In order for universities to provide staff for job market requirements, they should first be notified of those requirements. Based on these requirements, universities are those who draw up a curriculum suitable for these requirements, so that their students have easier employment after graduation. Our topic has a special contribution in this regard, since besides creating an automated model that will make a comparison between labor market demands and university curricula, we will also provide statistics and results on current requirements of the labor market in the field of technology.

3.10. Curricula suggestions

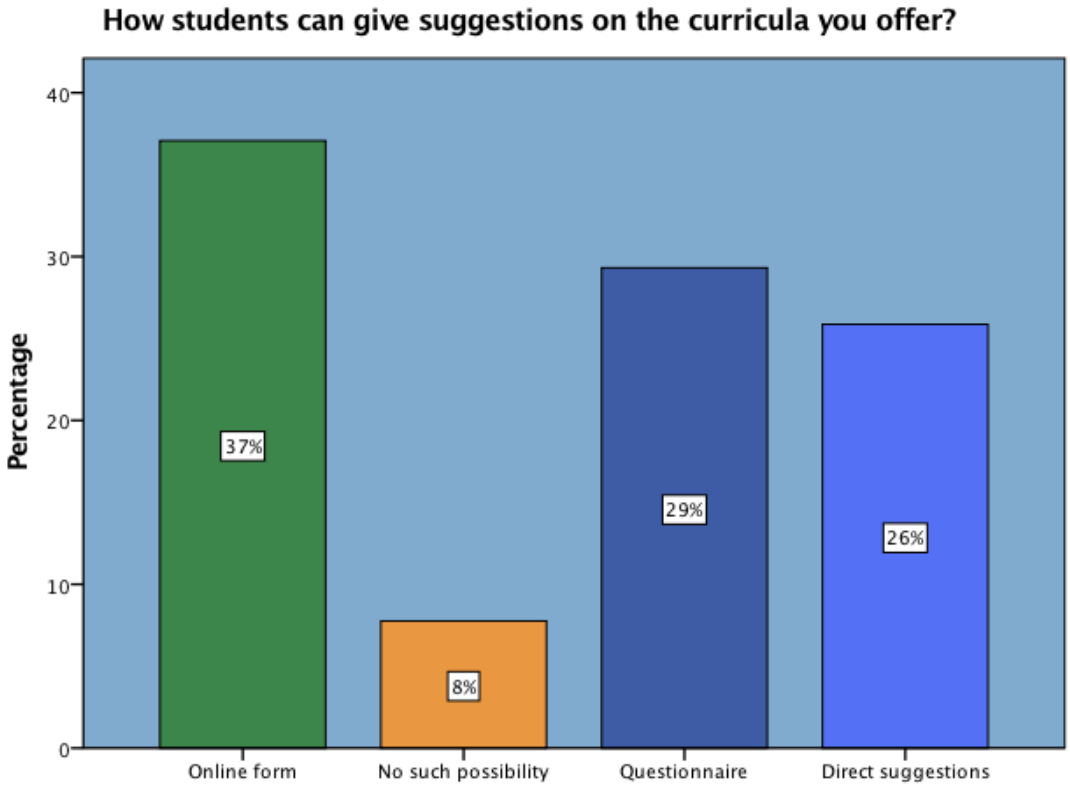


Figure 19. Curricula suggestions given from students

In Figure 19 we can see what are the ways students can give their suggestions on the curricula offered by universities. According to the processed results, it turns out that approximately 40% of them offer this opportunity to students through Online forms.

Approximately 30% of public and private universities offer this opportunity through questionnaires in physical form, and 26% prefer to get direct suggestions from students what should they include in the syllabus or not. A small percentage of 8% stated that their institutions do not provide such a possibility that students can make suggestions on the changes that the curricula should include.

Based on these suggestions that the universities take, the modification of the curricula they offer conforms to their study programs is also made. How often we modify curricula from universities in Kosovo, we will graphically present the processed results.

3.11. Curricula modification

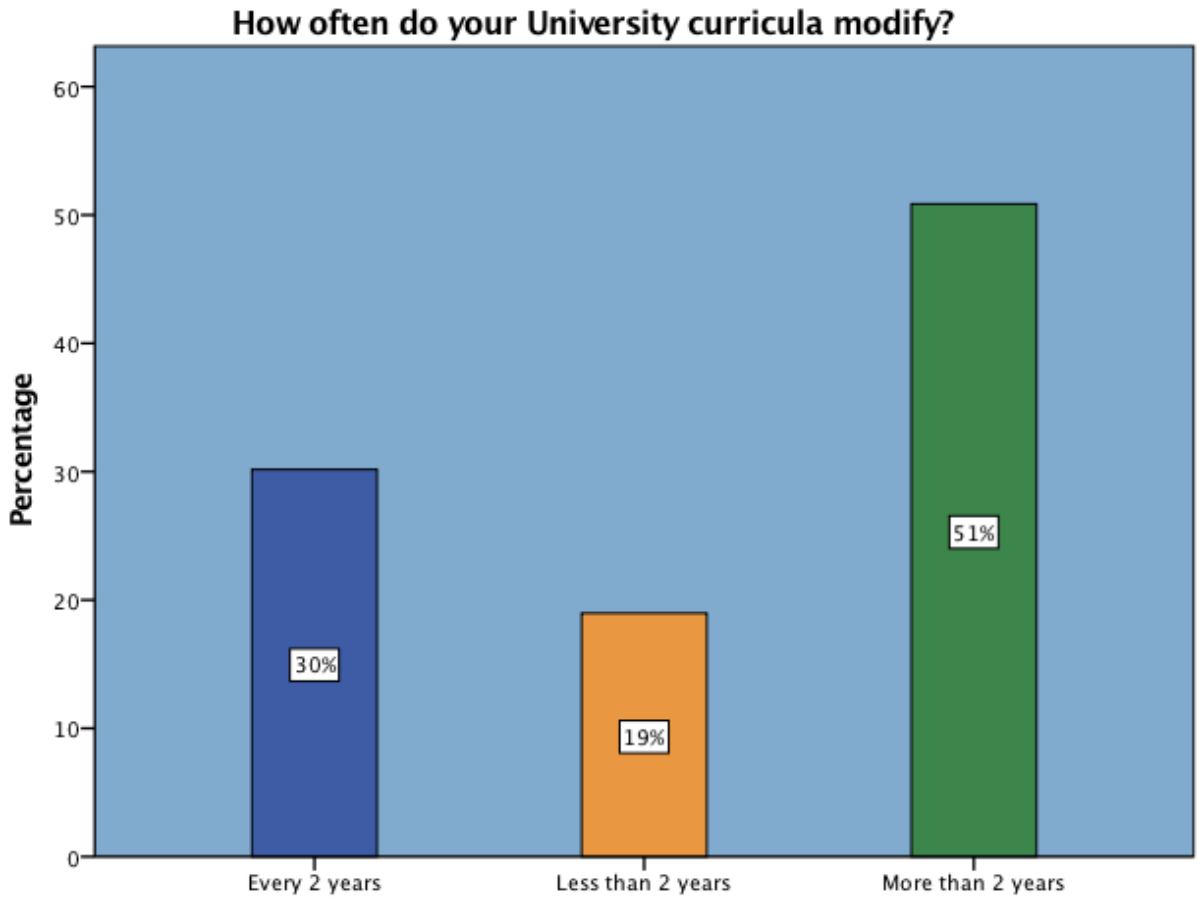


Figure 20. Curricula modification in universities in Kosovo

In Figure 20 we can see how often curricula are modified by public and private universities in Kosovo. From the processed results we can see that 51% of public and private universities modify the curricula at a time of more than 2 years.

Also 30% of them stated that they modified their curricula almost every 2 years, and approximately 20% of them stated that they modified their curricula in less than 2 years.

As we can see, curricula modification from universities in Kosovo takes place in a very long time, as 80% of them have been declared to modify at least every 2 years. In addition to modifying the curricula, we have tried to get results even on the application of any standard instruction on bachelor and master curricula. In the following, we will graphically illustrate how university students in Kosovo apply standard guidelines on designing their curricula that they offer in the field of technology. The guide has not been specified, but any guidance they use has been identified as usable.

3.12. Standard curriculum guidelines

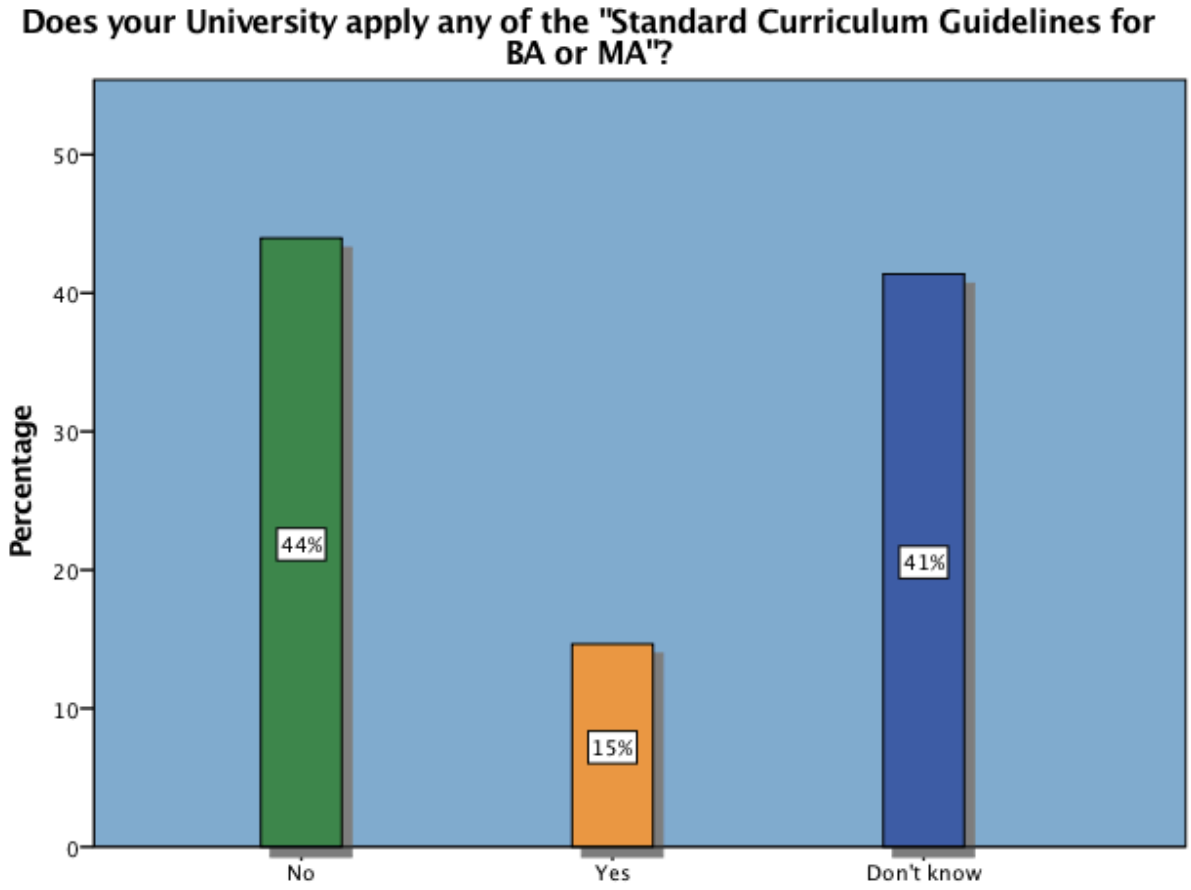


Figure 21. Standard curriculum guidelines for BA or MA in Kosovo

In Figure 21 we can see that a very small percentage has been declared applying standard guidelines for designing bachelor or master's curricula. Based on the processed results 85% of them have stated that they do not apply any standard guidelines for curriculum design. Of these, 44% stated that they did not apply one, and 41% of them stated that they did not know if their university applied to such a one or not. While a very small percentage of 15% stated that their universities apply such a guideline for the bachelor or master level.

Usually, the guidance provided by these curricula is of different nature, always influencing the raising of teaching and learning, for the only purpose of increasing the quality of the students who graduate in that university. According to these guidelines, the main goal is to prepare students for the job market, whether public or private. In the following, we graphically illustrate which sector is where graduate students in Kosovo are easily employed. So even at this point, the poll is made only with regard to the field of technology, since there are many technology companies operating in Kosovo.

3.13. Sector employment

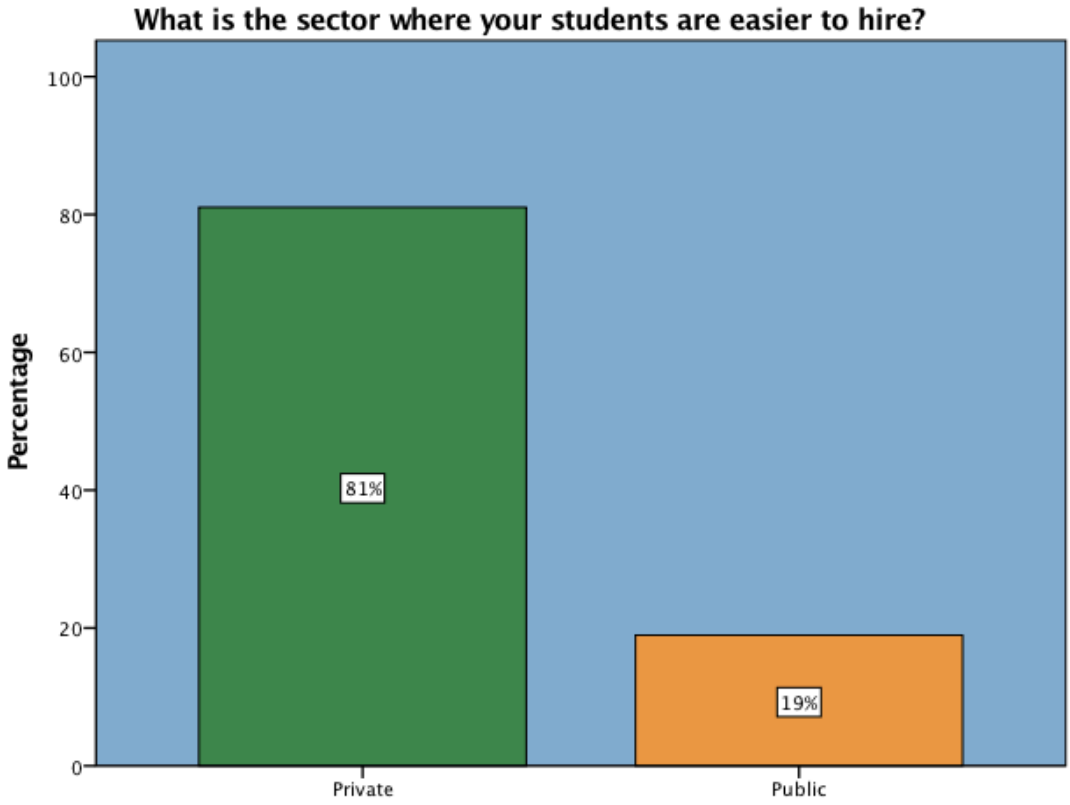


Figure 22. Private vs public sector employment

In Figure 22 we can see that the employment of graduates differs greatly from the private sector to the public one. Of all the universities surveyed, 81% of them stated that students have easier employment in the private sector, while 19% of them have stated that graduate students have easier employment in the public sector. There are several factors that have led to graduates being better employed in the private sector. Among these factors it is worth mentioning the agreement universities possess with different private companies for long-term student practices and after completing their studies. We have also conducted research in this regard, where we have received information on how many universities have taken over with private companies for different practices and what universities should include in the curriculum. This involvement of private companies in curriculum design is directly related to the fact that the students coming from that institution are more prepared for the requirements of that company, and there is no need for additional investments in the degree program to finish other trainings.

3.14. Cooperation between Universities and private companies

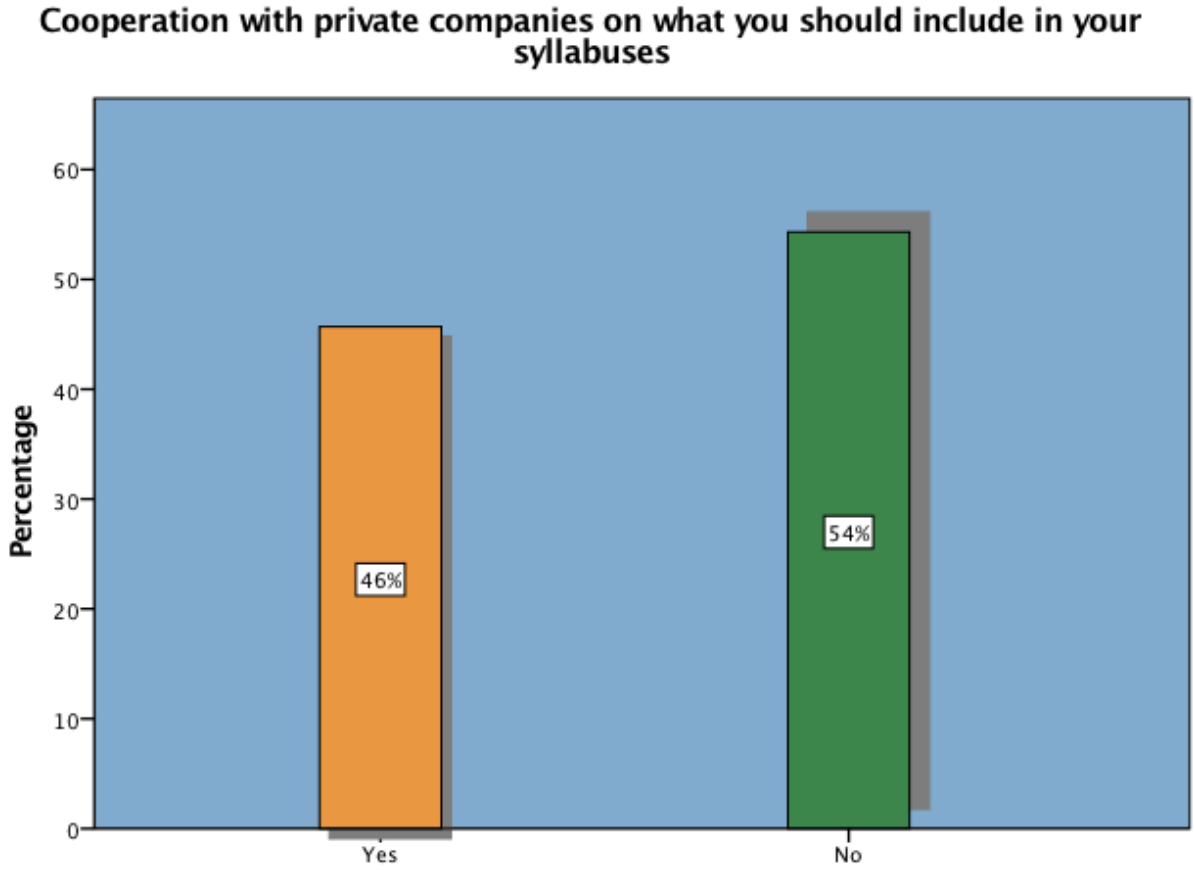


Figure 23. Cooperation between universities and private companies

In Figure 23 we can see how many universities in Kosovo have agreements with private companies in order to give suggestions what should be included in the curricula offered by universities. According to the processed results, 46% of them have said they have such agreements with private companies, while 54% of them have been declared not possessing such.

As we can see earlier, this may be one of the reasons why university graduates have easier employment in the private sector than in the public sector. Another factor is the fact that the number of private companies that offer technological services is much greater compared to the public sector. A disadvantage we are currently trying to identify in the private sector is the amount of net income that students can get after they get hired. Below we will graphically show what is the net amount of income the graduate students get to earn.

3.15. Students Net income

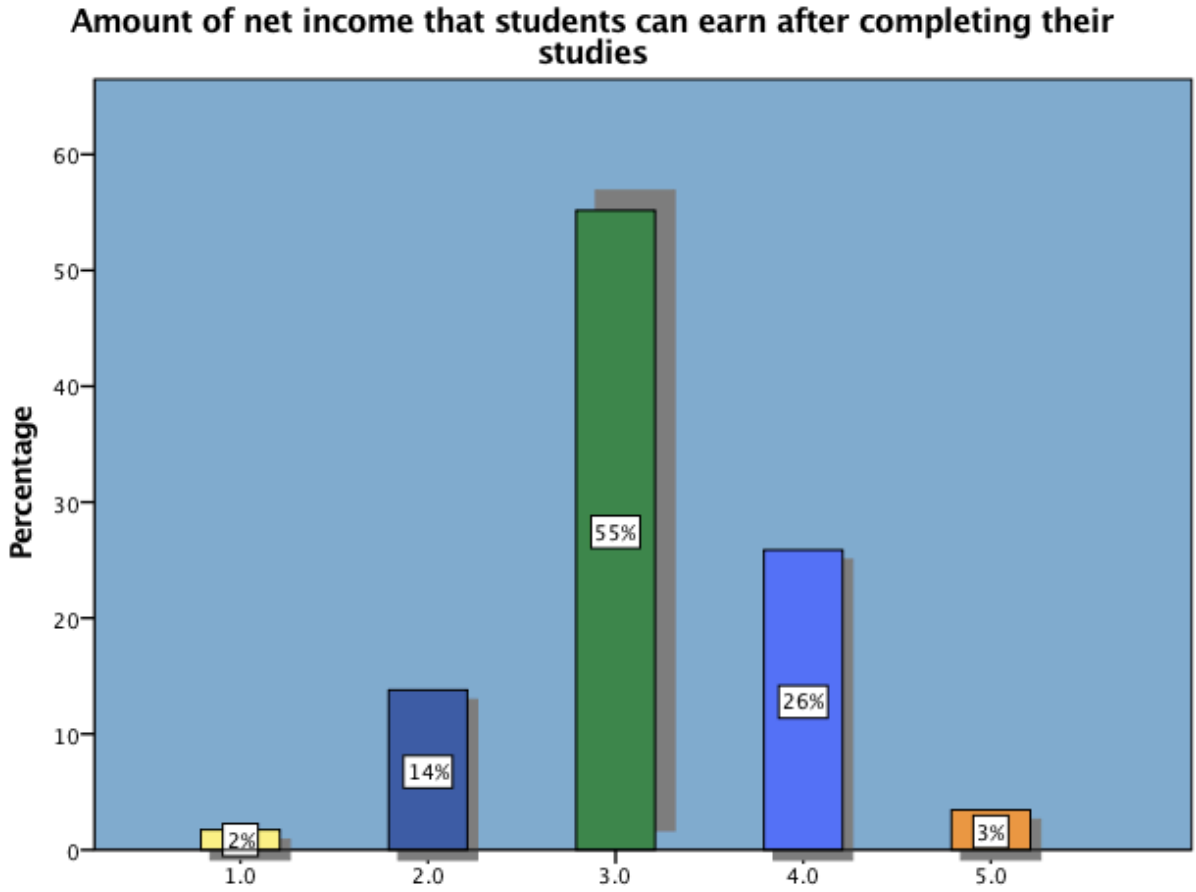


Figure 24. Students net income after completing their studies

In figure 24 we can see that net earnings of students who get to work after graduation are an average level. It is the private companies that have implemented knowledge in the fields: *programming, databases, computer networks*, etc. to create average income compared to other areas that exist in Kosovo.

However, there are shortcomings in the preparation of students who are graduating in the field of technology, as most of them are forced to complete different practices and trainings in order to be ready for the labor market requirements that exist in the field of technology.

This inconsistency of labor market demand with university curricula also has the effect that many students work for jobs that do not correspond to the field of study they have completed. Our research has been extended in this regard as we have tried to get results from universities rather than the work the students make regarding the field in which they graduated.

3.16. Current work of students

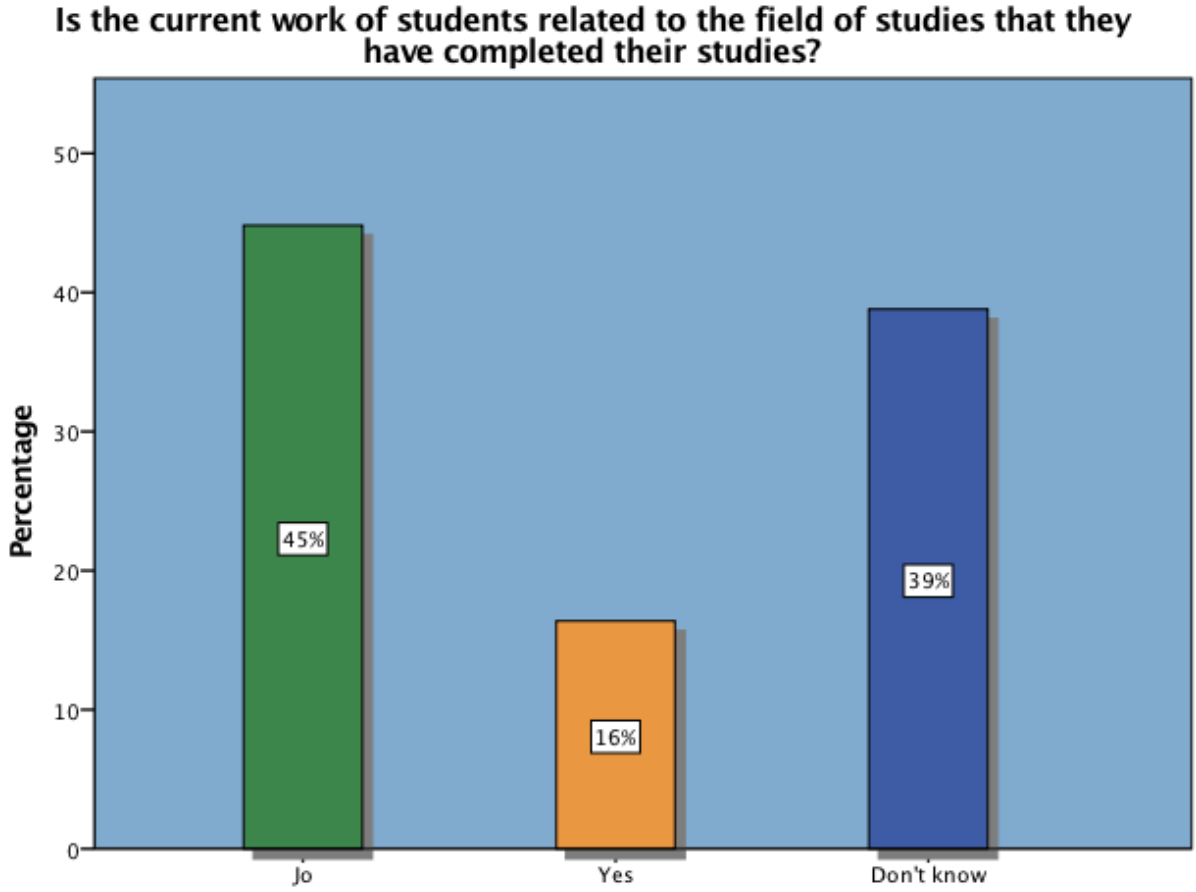


Figure 25. Current work of students vs field of studies

Completing the studies in the right direction is of great importance to both the student and the university. Based on the results that have been processed by 45% of them, it is stated that the graduate students do not work which corresponds to the field in which they have completed the studies.

Also 39% of them stated that they did not know if their students were working in the same field where they graduated, and 16% of them stated that students who graduated from their universities worked the same field they did have completed their studies.

Of great importance is given to work in the same field with that of studies in other fields, as according to the literature is one of the key factors that affects the motivation of the work, resulting directly in the best performance of the worker in that field. It is precisely our research that will directly affect that graduate students can work in the field in which they have completed their studies. Looking at the importance of such a model we will present how many universities think this model is important.

3.17. Automated model importance

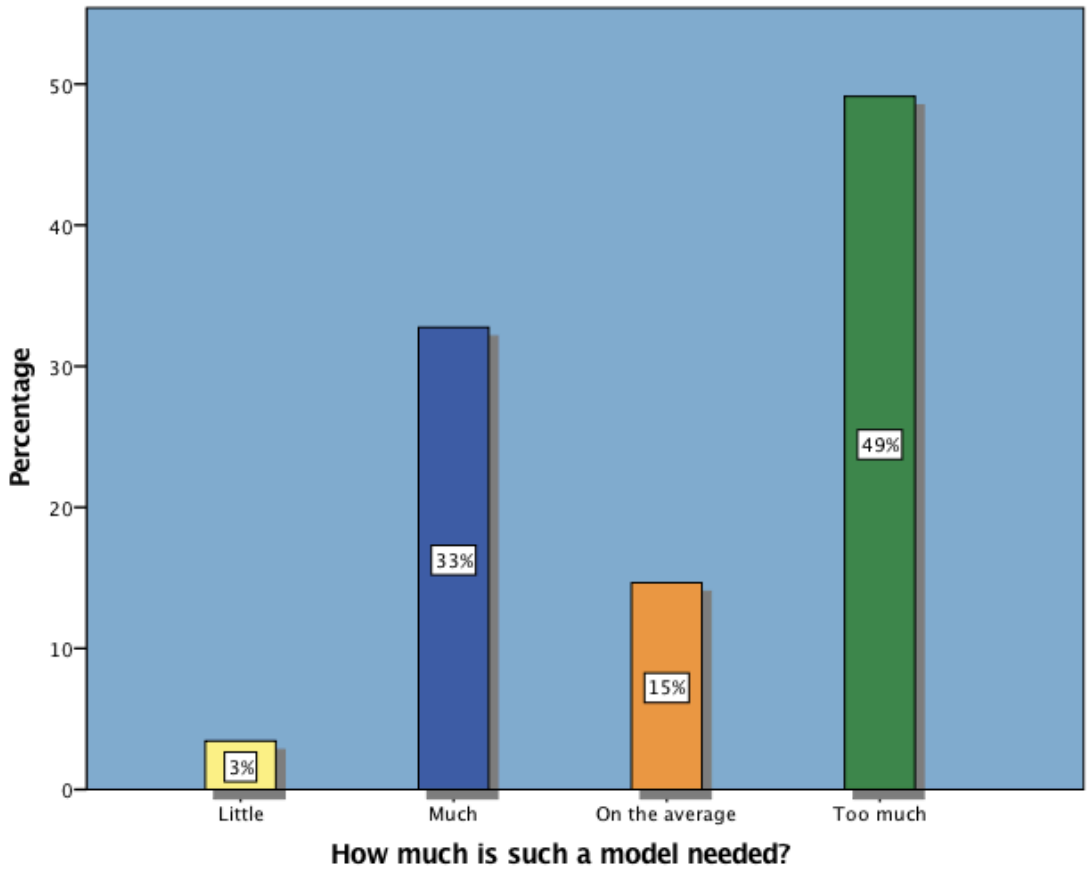


Figure 26. Automated model importance

It is also anticipated at the outset that great importance is given to the creation of such a model that will make a comparison between labor market demands and curricula offered by universities. Based on the results produced by public and private universities in Kosovo, approximately 50% of them think that such an automated model is needed at a high level. Also 33% of them think that such a pattern is needed above the average level, and 15% think that such a pattern is moderately necessary, and a very small percentage of 3% stated that a model I this is a little needed. Part of this survey has been the university's readiness to contribute directly to the implementation of such a model in various forms, such as the application of such a model after implementation and the provision of access to the data the universities possess.

Next we show results of private companies in the field of technology in order to see how necessary is such a model.

3.18. Area of employer's bachelor's degree

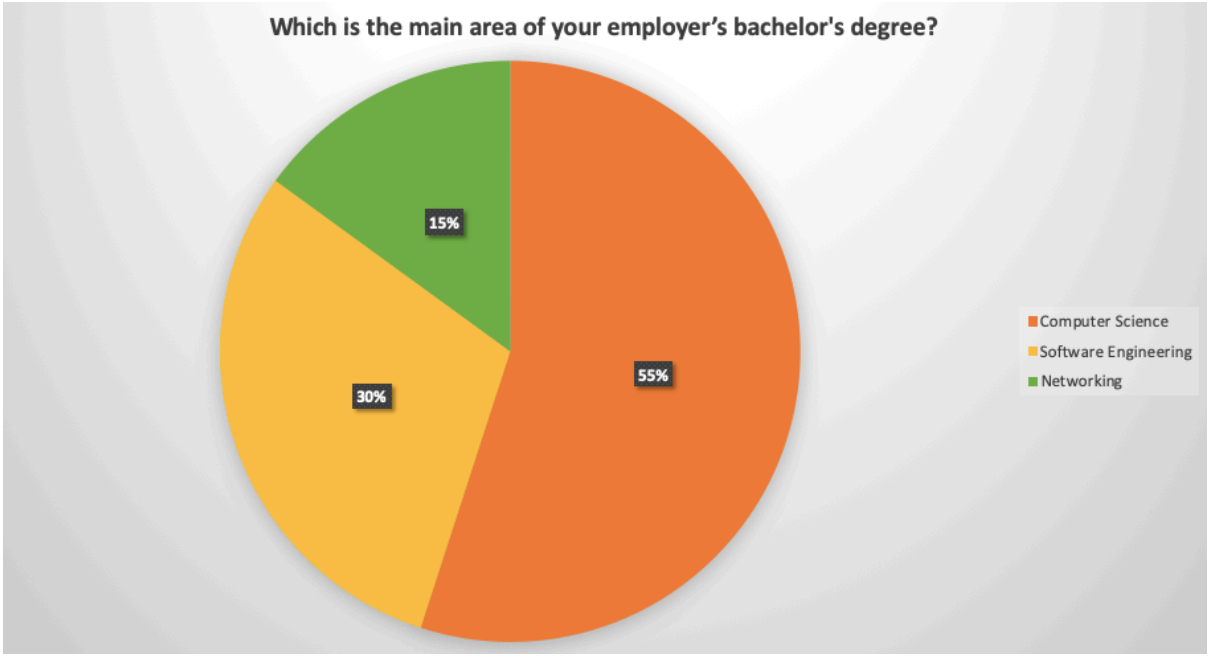


Figure 27. Area of employer's bachelor's degree

In figure 27 we can see the area of employer's degree in bachelor level. As it is shown in figure 27, there are three areas in the field of technology that employers are employed. The area of Computer Science covers 55% of employers, the area of Software Engineering covers 30% of employers, and the Networking field covers only 15% of employers.

We can conclude that the focus of students in the field of technology is mostly separated in Computer Science area and Software Engineering area. Next we show what is the degree that employer's own.

3.19. Title of employer's bachelor's degree program

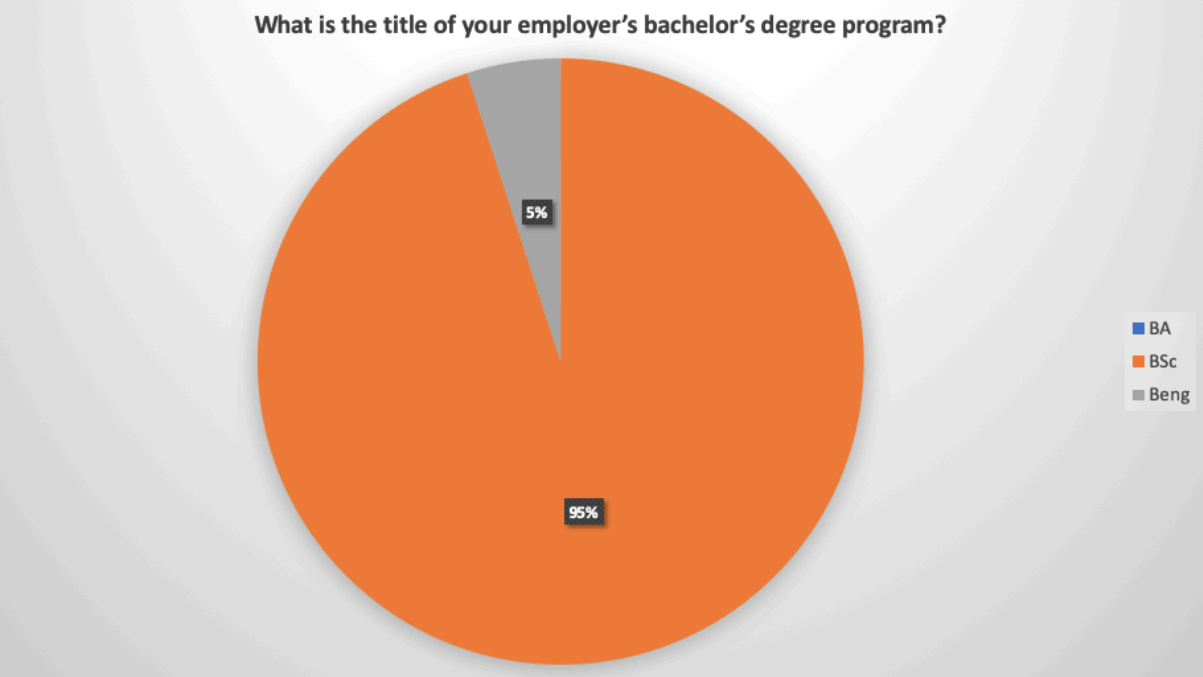


Figure 28. Title of employer's bachelor's degree program

Figure 28 shows the title of employees in private companies in the field of technology. As we can see, the largest number of tech workers is with the BSc call where we have 95% of them, while the BEng call has 5%. These differences appear in the education system that students have previously completed as a four-year system. In the following we will show how many of the employees have continued their postgraduate studies and received the master's degree. Knowing the importance of master studies, later we will present even more in-depth analysis of whether students who have completed postgraduate studies can solve problems that appear in the company after their employment.

Further analysis will also be done on which areas are best covered by students who have completed these studies.

3.20. Employers master degree

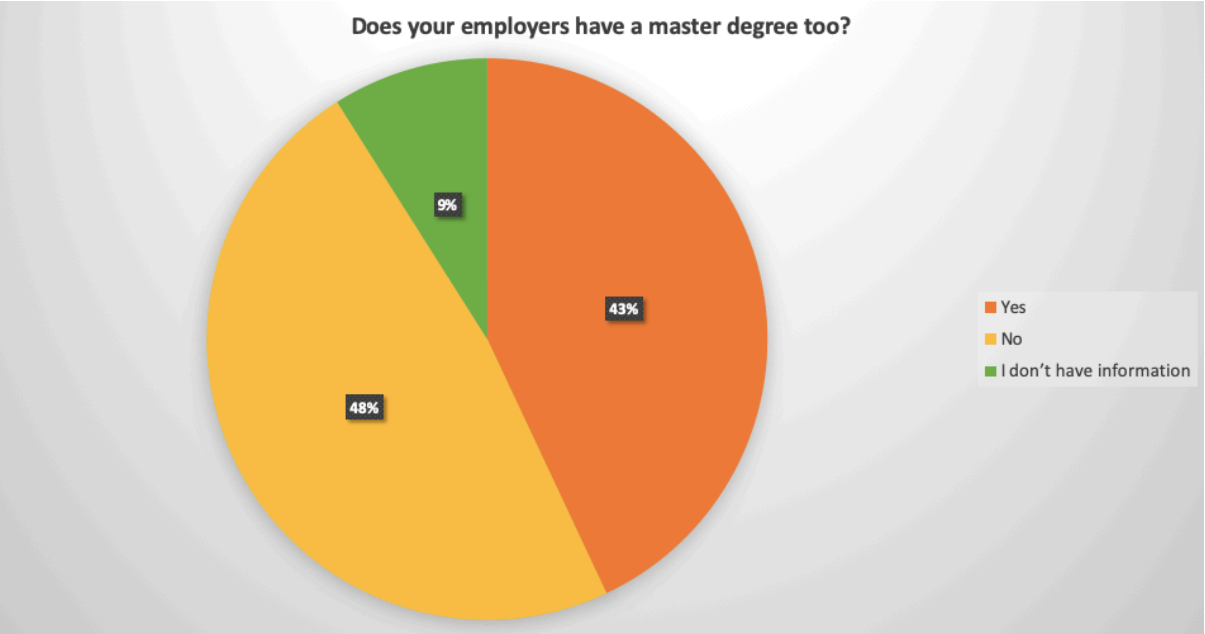


Figure 29. Employers master degree

As can be seen in figure 29, the number of employees who have completed master studies is less than those who have not completed their master's degree. The number of students who have completed master's degree is 43%, while the number of students who have not completed master's degree is 48%. Of these, 9% stated that they did not have information on whether their employees had completed these studies or not.

According to the statements made by company leaders, it is precisely the small knowledge that students achieve during their postgraduate studies that is why they do not pursue a master's degree.

Certainly, a better adaptation of technology syllabuses in the field of technology would have a positive impact in this regard, making students who are employed in various companies able to solve the problems that arise. Of course, in other analyzes we will also show how young hires are able to solve the problems that arise in their jobs and whether these problems are obstacles to the development of companies.

In the following we will present what are the main areas in which employees have completed their master studies.

3.21. Area of employer's master's degree

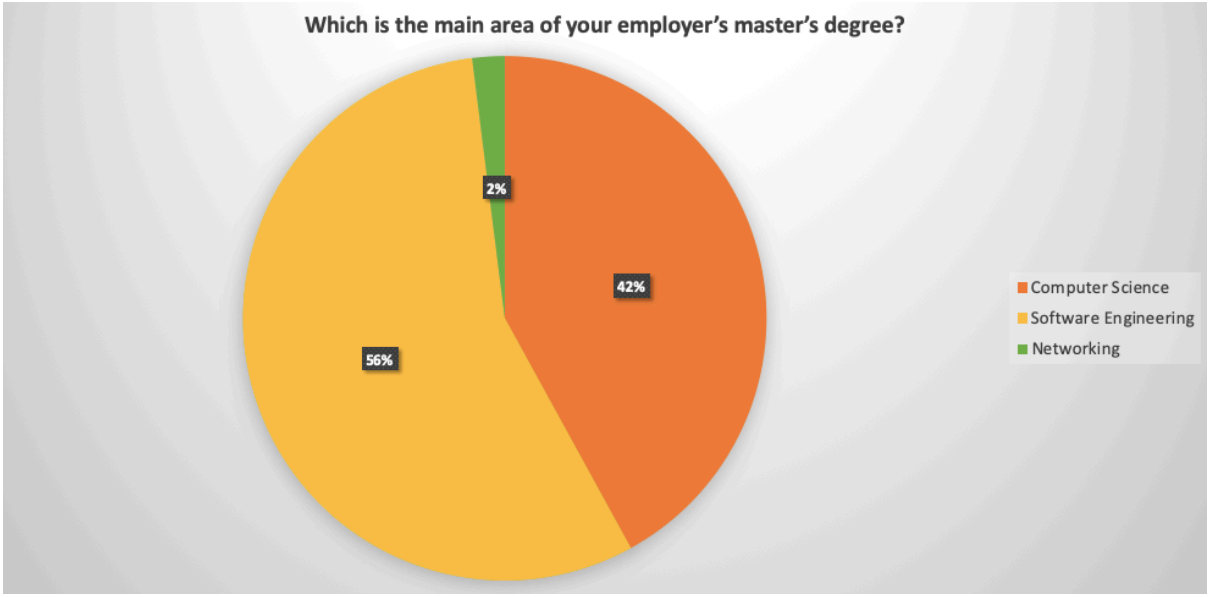


Figure 30. Area of employer's master's degree

As shown in figure 30, the largest percentage of the field in which students have completed master's degrees are Computer Science and Software Engineering. Of the companies that responded to this questionnaire, 56% stated that the field in which they completed their master's degree is Software Engineering, while 42% stated that the field in which they completed their master's degree is Computer Science. A small number of 2% stated that they have completed master studies in the field of Networking. Of course, such a small percentage of employees in the field of Networking indicate that the requirements are greater in this field and that it is required to offer study programs in the field of Networking.

In the following we will present the analysis of the title of the employees who have completed their master studies. Of course, this also depends on the system of studies that employees have completed and the place where they have completed.

3.22. Title of employer's master's degree

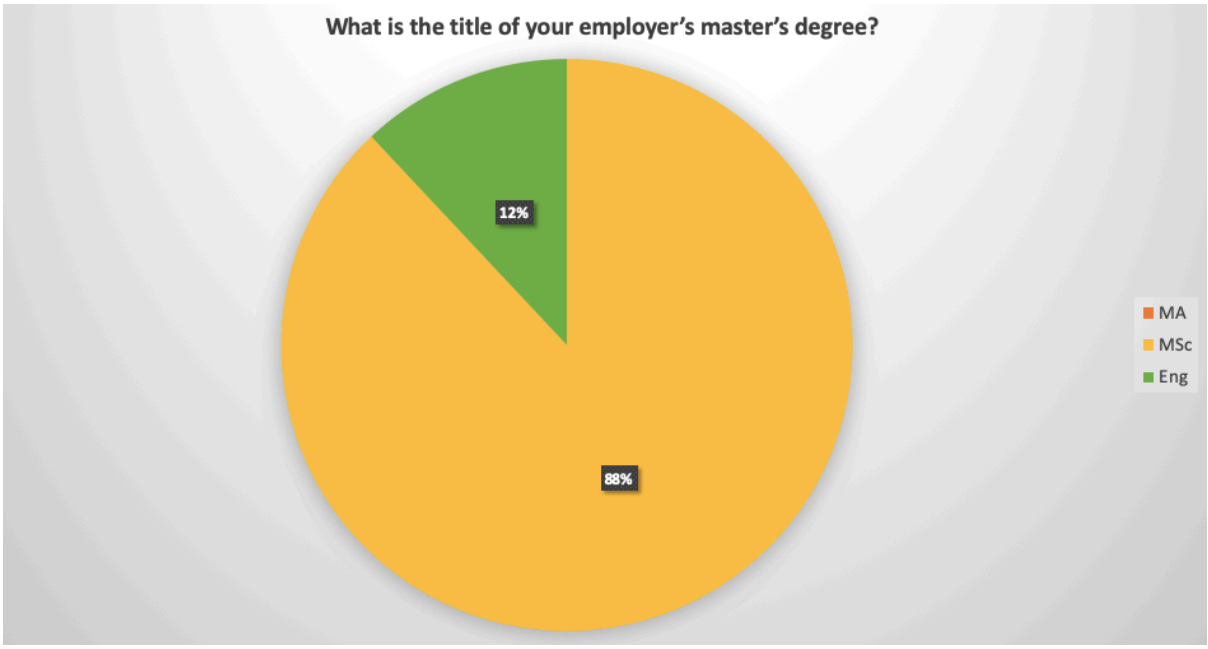


Figure 31. Title of employer's master's degree

The title that employees have in private companies in the technology field is shown in figure 31. As you can see the largest percentage is in the title of master of science where 88% of them stated that they have completed their studies there. system. While 12% of them stated that they have the title of engineer, a title that depends directly on the system of studies that have been completed. Below we will present what are the positions offered in some of the companies that have completed the questionnaire.

3.23. Available position jobs in companies

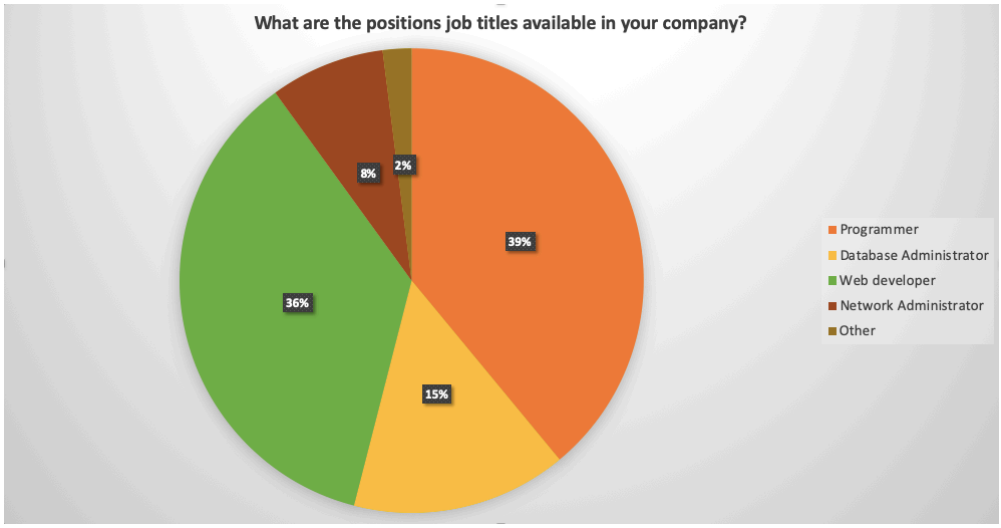


Figure 32. Title of employer's master's degree

In figure 32 we can see what are the positions offered in the companies that have completed our questionnaire. As we can see the positions offered by companies in the field of technology are: Programmer, Database Administrator, Web developer, Network Administrator and others.

Of all the positions we have mentioned the largest number of positions offered by private companies in the field of technology is Programmer and Web developer. Of all companies, 36% of them stated that the vacancies offered by the company are Web developers while 39% of them stated that the vacancies offered by the company are Programmers. This allows us to understand that the degree programs offered by universities are required to provide as much knowledge as possible in these two fields as they are the areas with the most technology vacancies.

Compared to these two positions, even for the Database Administrator position we can conclude that there are vacancies, as 15% of all companies have stated that they offer positions in this regard. Also for this position we can conclude that universities are required to provide as much knowledge as possible to their students in order for them to be able to solve problems when they reach employment in a company.

Of all of these, a small number have stated that they are offering vacancies in the field of Networking, with only 8% of all companies claiming to be in the field. While only 2% of all companies stated that they offer jobs in other fields not mentioned in our questionnaire.

Such research is of great importance since we have so far had no statistical data which has shown the labor market requirements offered in our country.

In the following we will present an analysis that has been done to companies on how much employees are able to transfer the skills acquired from bachelor degrees to the companies where they are employed.

Evaluation is done by company leaders who have stated what are the skills employees have when they reach employment after completing their bachelor's degree.

3.24. Employer's skills gained from bachelor studies

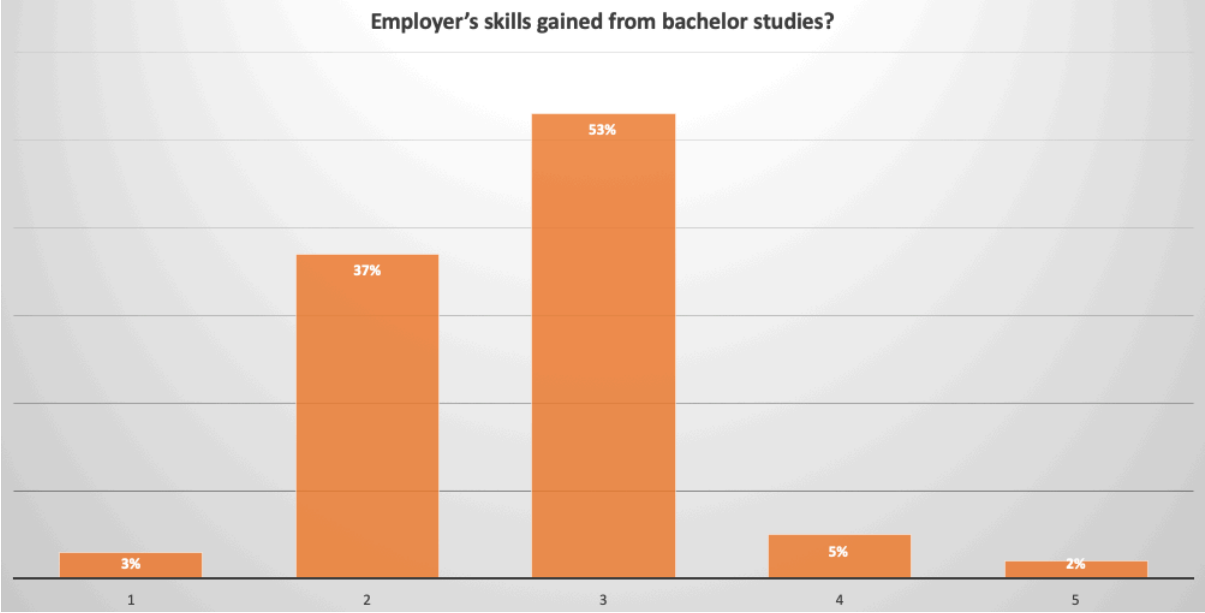


Figure 33. Employer's skills gained from bachelor studies

Figure 33 shows the company leaders' assessment of the skills their employees acquire after completing their bachelor's degree. The rating has been from 1 to 5, where 1 represents the lowest level of knowledge as well as the 5 highest level of knowledge gained during bachelor studies.

According to the answers given by company leaders, the knowledge that their employees have is at an average level of 2 and 3. Only 3% of companies responded with the lowest rating, while 37% of companies responded with 2. .

The average rating which was number 3 was 53% of the companies, while the highest rating was lower.

5% of the companies answered for Level 4, while only 2% answered for the highest level in the questionnaire.

From this analysis we can conclude that higher knowledge is required during bachelor studies in order for employees to be able to transfer this knowledge to the companies in which they are employed. In order to analyze whether these results are consistent with a subsequent analysis, we will present an analysis of whether employees are prepared for the job they are doing.

3.25. Employers prepare for their current job

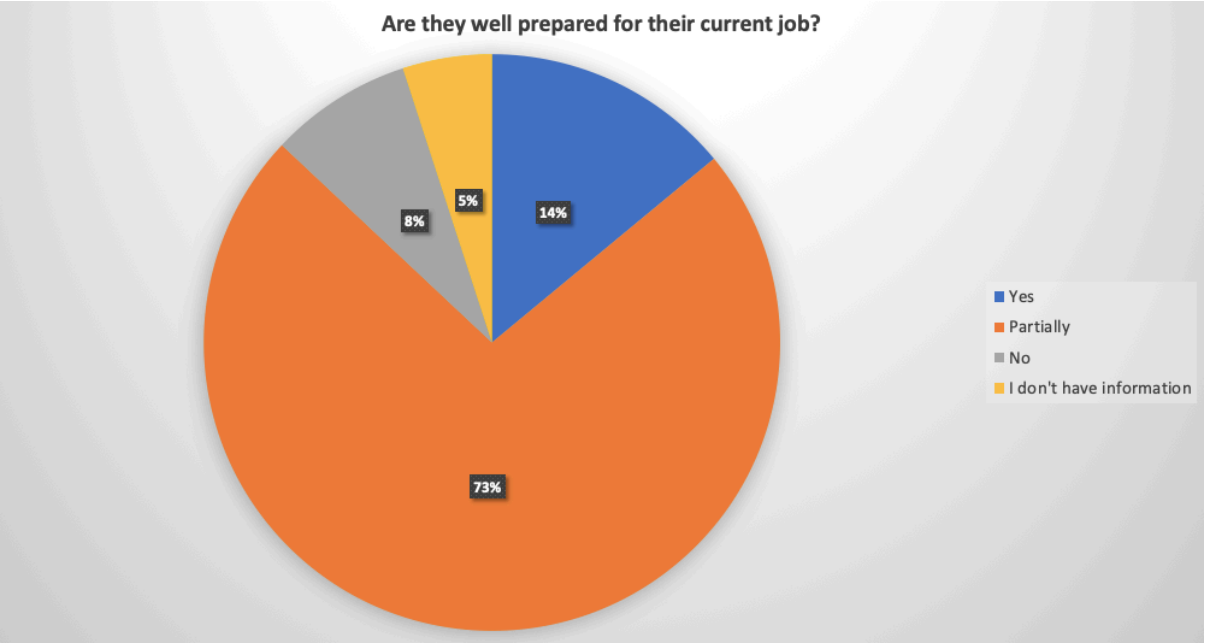


Figure 34. Employers prepare for their current job

Figure 34 shows the analysis of how well employees are prepared for the positions they hold in private technology companies.

According to the answers given by the company leaders, it turns out that 14% of them are prepared for the job they hold. Of these, 73% are partially prepared which is a satisfactory average. Of these, 8% stated that their employees were not prepared for the positions they hold in their companies, and 5% stated that they did not have such information.

Based on these results we can certainly conclude that there is still a lack of preparation of the students with whom they come out after completing their bachelor studies. In order to analyze whether there is a correlation between these two answers we will make an analysis.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	58.319[a]	9	2.823E-9
Likelihood Ratio	53.807	9	.000
N of Valid Cases	91		

Figure 35. Student preparation for current job variance

Figure 35 presents the Chi-Square test of 91 responses from all companies, and it turns out that we have a very high dependency between the preparation with which students come out after the end of their bachelor studies and the knowledge they have about the work for which

they keep it. We say that we have an addiction between them since the signi ne cance between them is much smaller than α which should be less than 0.005. In the following we will show how companies think that university education is balanced between the theory that students take and the practice that students should implement after their studies.

3.26. University balance between theory and practice

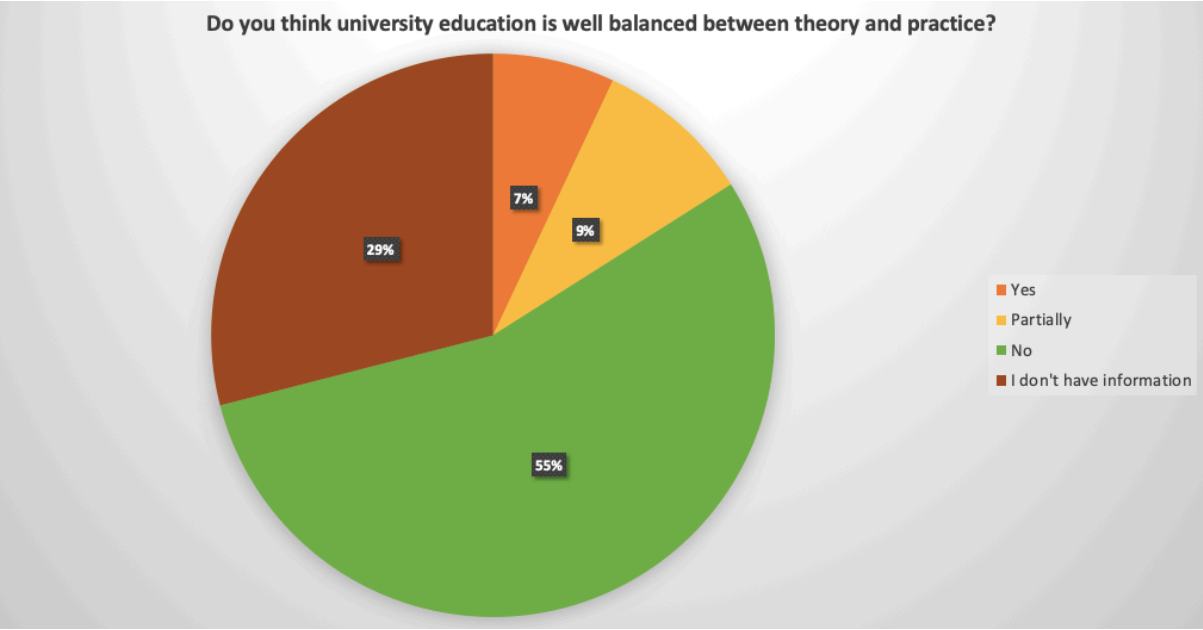


Figure 36. University balance between theory and practice

Figure 36 shows how companies think university education is balanced between the theory that students learn during their studies and how well they are able to apply that knowledge. As we can see, 55% of them stated that there is no balance between the theory that students acquire and the practice of that knowledge. Of these, 7% stated that there is a balance between theory and practice from university education, 9% stated that there is partially a balance between theory and practice from university education. As well as 29% stated that they do not possess such information. These data show that students during bachelor's and master's degrees are required to do more internships in order to be able to practice what they learn during the teaching process. In the following we will present an analysis of whether more electives should be included in the study programs.

3.27. Elective courses in university curricula

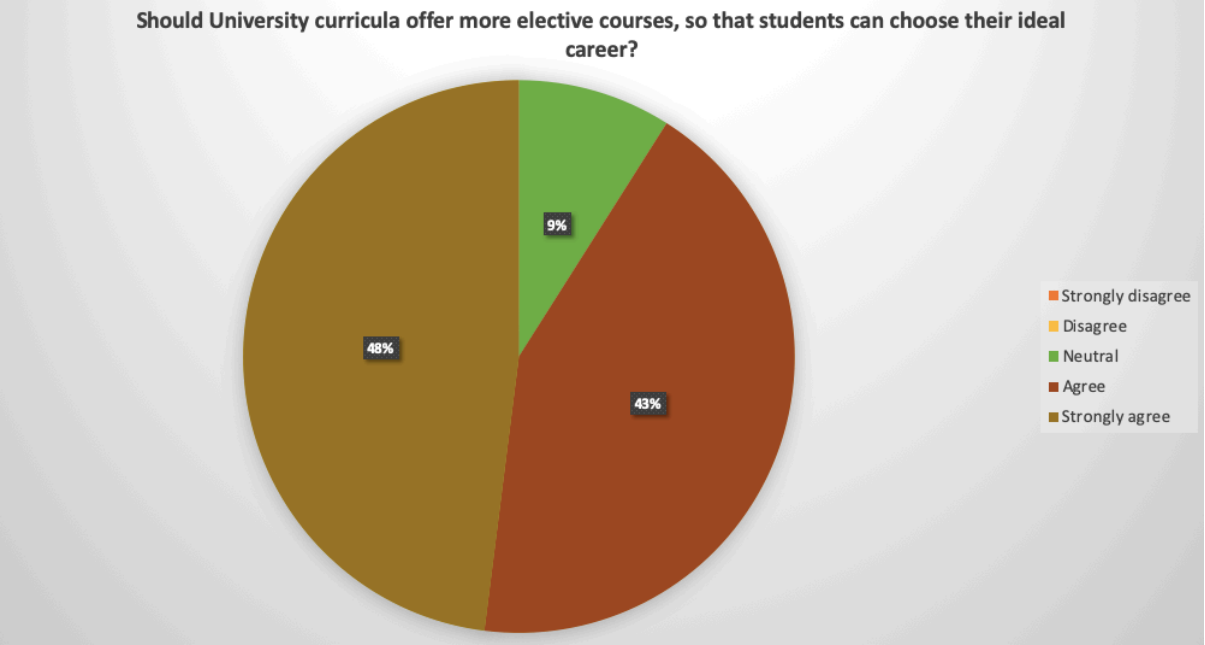


Figure 37. Elective courses in university curricula

The importance of electives in studies is shown by the fact that students are able to transfer that knowledge later to the companies they employ. In order for students to have greater opportunities to choose their ideal position, more elective courses are required during their studies. What private companies think about this issue is shown in figure 37.

As we can see companies have been able to respond strongly disagree, disagree, neutral, agree and strongly agree.

Of all the companies, none of them think there should be electives included in the study programs offered at universities. All responses are neutral, agree and strongly agree. 48% of them strongly agree, 43% agree while only 9% are neutral.

Such an analysis helps us to conclude that universities should include as many electives as possible in their programs in order for students to be able to choose their ideal position after completing their studies. In the following we will present an analysis of whether students are able to solve problems with knowledge coming from the university.

3.28. Problem solving by students

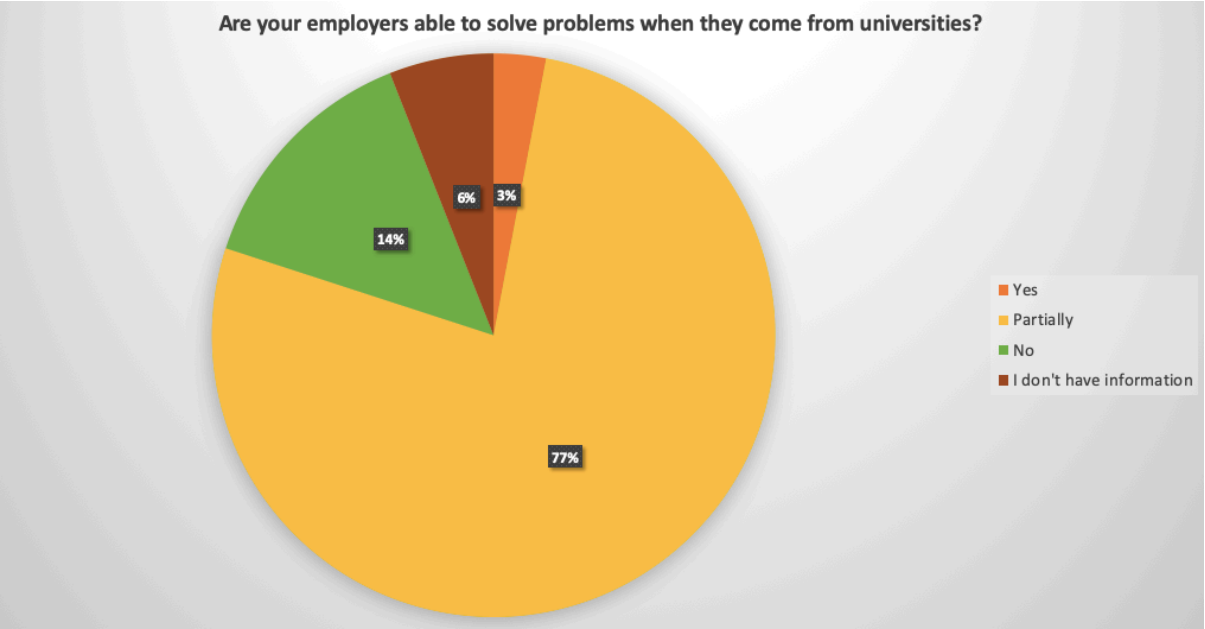


Figure 38. Problem solving by students

Problem solving in the company is of great importance for the position of the employee. Of course, these problems should be able to be resolved by the employees based on the knowledge they have acquired during their studies. Figure 38 shows how students are able to solve problems when they come from university. According to statistics, only 3% of them are able to solve problems, 77% stated that they are partially able to solve problems, 14% stated that they are unable to solve these problems, while only 6% stated that they do not possess them. such information.

One factor that influences these problems to be solved more easily by workers is the practice that they manage to do during their bachelor or master studies, so that those practices can be transferred to companies as they are hired. . Such a statistic will certainly help us to have more accurate information about employees so that we can make recommendations for study programs offered by universities in the field of technology. To support this analysis, we will present the next analysis which shows which areas need to be better covered by the region's universities.

3.29. Areas that need to be covered from Universities

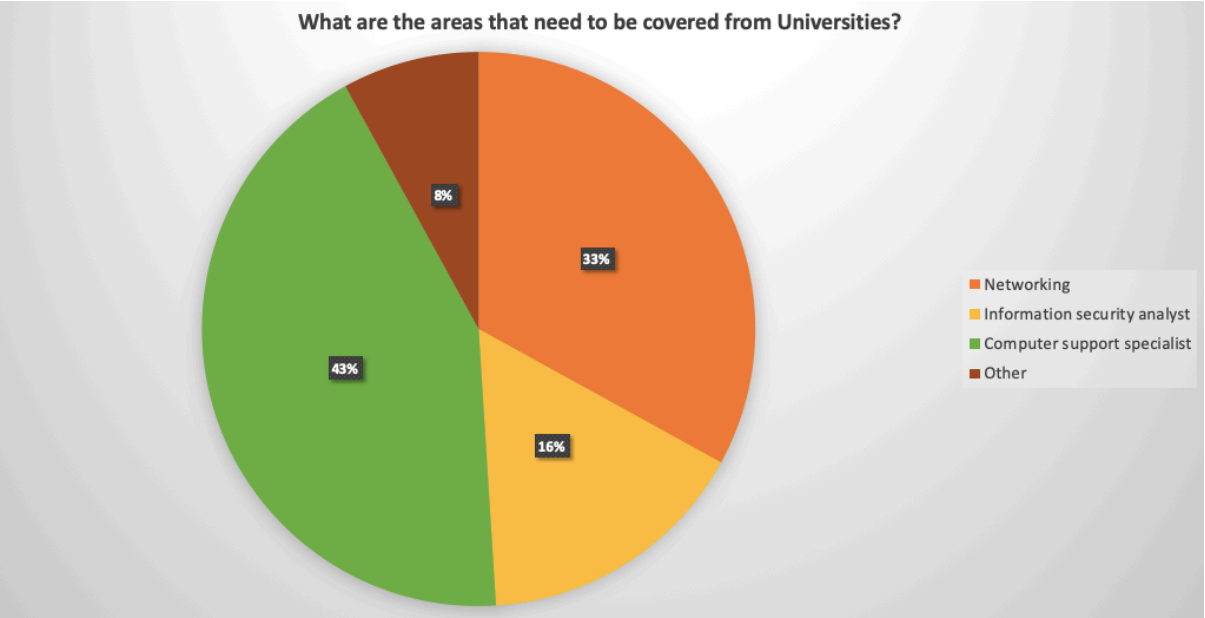


Figure 39. Areas that need to be covered from Universities

Labor market demands show that there is an increase in technology jobs. As presented at the beginning of the topic, an analysis shows that after 2020, there will be 8.1 million technology jobs. Such a statistic is too big to be completed by technology graduates. This is also the reason that has forced us to analyze the situation which are the areas that need to be better covered by universities, so that after 2020 we will be able to provide frameworks for the regional market but also for the European one.

Based on our analysis there are three areas that need to be covered more by universities in order to meet the demands of the labor market in our country. According to private companies 43% of them stated that the area which should be better covered by universities is computer support specialist. This field is not accidentally chosen, since every business, even of an economic nature, is directly dependent on technology. The second area that needs to be better covered by universities is Networking, where 33% of all companies stated that this area needs to be covered. Of those, 16% stated that the area which should be better covered by universities is Information security analyst. While 8% of them stated that they are other areas that should be covered besides the ones mentioned above.

In the following we will present the analysis that shows whether private companies offer training programs for new employees coming to their companies.

3.30. Training program for new employees

Figure 40. Training program for new employees

The preparation of employees can also sometimes depend on the company in which they are employed. Of course, this can be achieved through special training programs provided by the company. The company may decide to offer training programs on its own, but it can also tell employees abroad to train them in specific areas.

Figure 40 shows the analysis of whether such programs are offered by private companies in the region. Based on our analysis, 98% of all companies stated that they do not offer such a program and do not send employees abroad to be treated in certain areas. Only 2% of them stated that they offer such an opportunity for their employees by training new workers or sending them abroad. Of course, such a statistic is worrying given the fact that student preparation is still directly dependent on the universities in which they graduate. Since most companies have stated that their employees do not send in training or provide training programs for them, we have analyzed what are the areas in which employees should be sent to training.

3.31. Areas that employers need to be trained

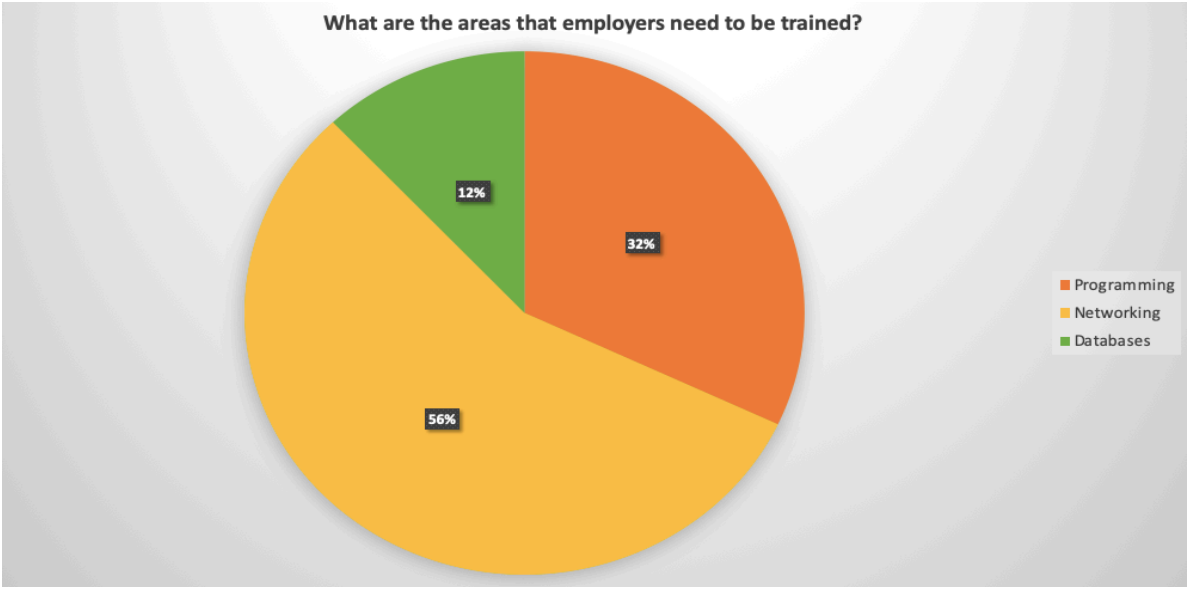


Figure 41. Areas that employers need to be trained

In figure 41 we show what are the three main areas that workers need to be trained. The top three areas to choose from are Programming, Networking and Databases. Of these areas, 56% of companies stated that the area that needs workers to be treated after coming from universities is Networking. For the Programming field 32% of them declared, while for Databases 12% of the companies stated.

Even this statistic will be of great help to universities and the market, as through this we can know which are the areas to be incorporated within the study programs. In the following we will present the Chi-Square Test between this analysis and the level of preparation with which students come from the university in order to see if there is any dependency between them.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	54.449[a]	12	0.000003
Likelihood Ratio	54.479	12	.000
N of Valid Cases	91		

a. 12 cells (60.0%) have expected count less than 5. The minimum expected count is .18.

Figure 42. Areas that employers need to be trained variance

Figure 42 shows the Chi-Square Test which shows that there is a great deal of dependency between the areas that students should be treated and the preparation with which they leave

university. As can be seen in Figure 42, the signi sesa cance is much lower than the α = 0.005 level.

The following will present an assessment of how well workers are prepared for practical or industrial work.

3.32. Employer's industrial or practical preparation in company

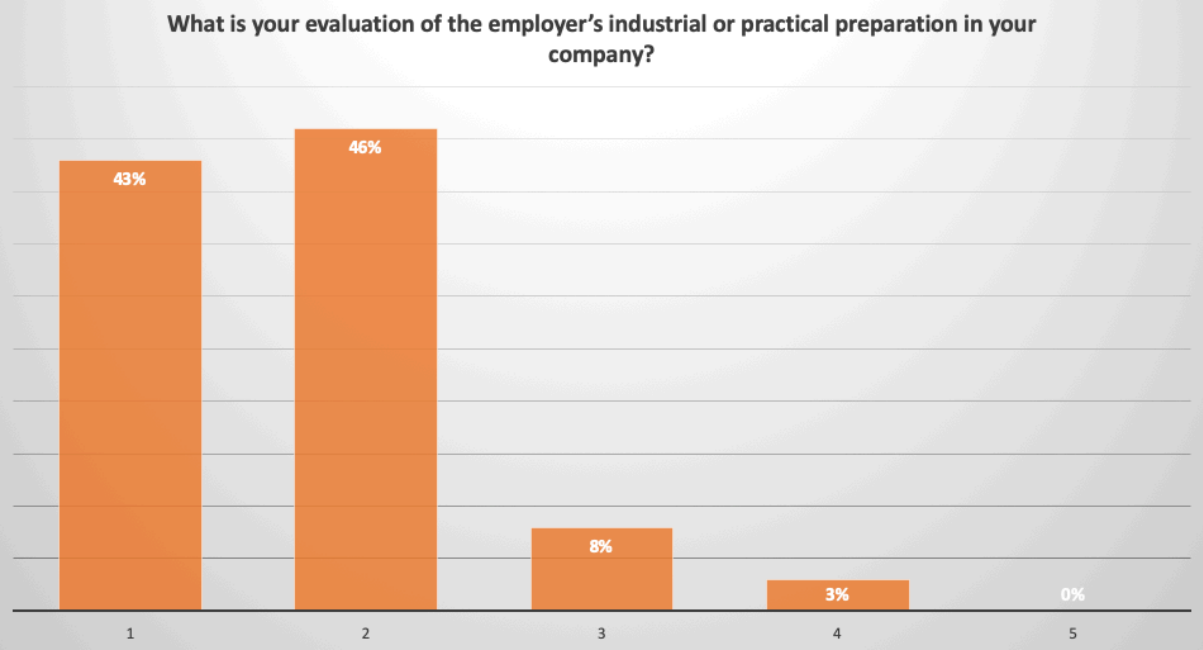


Figure 43. Employer's industrial or practical preparation in company

Figure 43 shows the ratings made by companies about industrial or professional training for their employees. Valuation is made of value 1 as the lowest value, and value 5 as the highest value. According to the companies, 43% of them rated the lowest value of their employees' industrial or professional training. While 46% of them rated it 2, it is still at an unsatisfactory level of preparation. On average, only 8% of them rated it, and only 3% of companies rated it above average. While the highest value was not answered by any of the companies that were part of our questionnaire.

Even this analysis shows that companies are not satisfied with the professional or industrial training that workers possess after graduation. This is even more supportive of the fact that study programs offered by universities in the region should include as much internship as possible for graduating students to be as prepared for the regional and European market as possible.

3.33. Job vacancies announce form

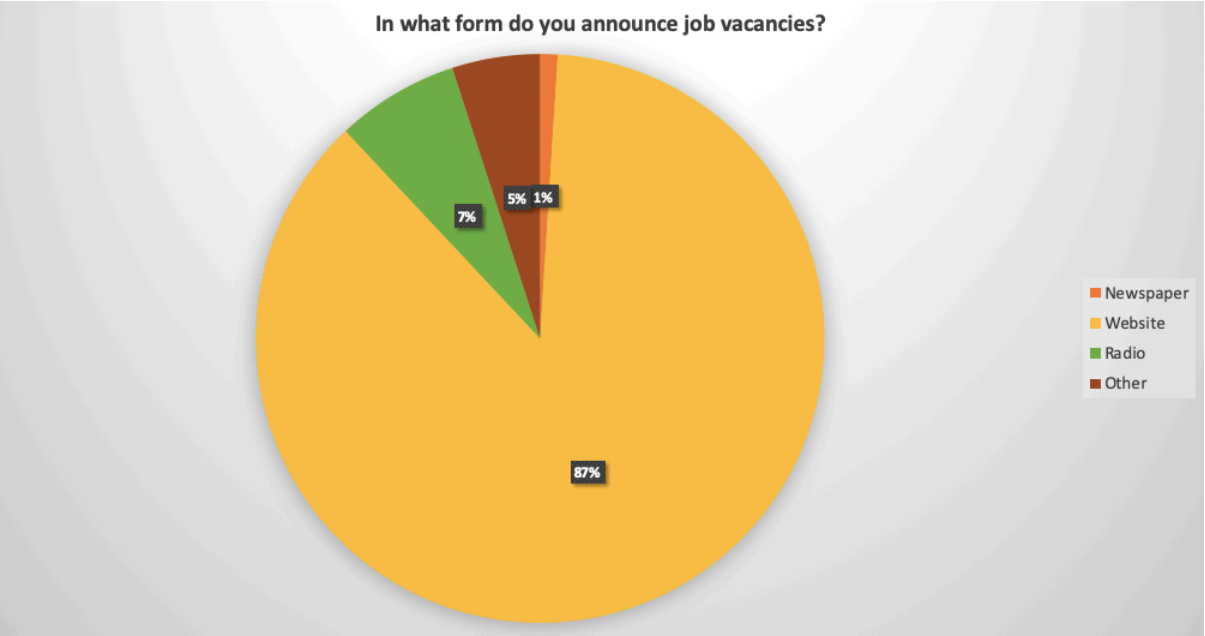


Figure 44. Job vacancies announce form

Figure 44 is a form of how private companies publish information about vacancies that are free in their companies. In the questionnaire which is distributed in the company, there were several possibilities for them to answer how they advertise the vacancies when there are vacancies in their companies. How likely they were to respond: newspaper, website, radio and other. And based on responses from private companies, 87% of them post ads on websites that publish job offers. Of all companies, 7% of advertisements advertise on the radio, 1% advertise in newspapers, and 5% of them advertise advertisements in other forms. Based on this research, this motivates us even more for our research to be conducted in this direction, as we will extract data from the websites that publish them. Considering that almost all companies publish these competitions on the website, then it is our model that will automatically extract information from these websites. In the analysis taken by students and academic staff, they also responded that they receive almost all notices from websites publishing competitions in the field of technology.

In the following graph we will present the analysis of the impediments in the development of the company, the preparation of the students coming from the university.

3.34. Development of companies

Figure 45. Development of companies

In figure 45 we present the results obtained from private companies in the region, where they answered whether the level of student preparation is an obstacle to the development of their company. The answers that could have been obtained from the companies were: yes, partially, no and I have no information.

Based on the results obtained from companies, 58% of them responded that partially disadvantaged students are an obstacle to the development of their company. Of these, 34% responded directly that student non-preparation is an obstacle to their development. Of all companies, 7% of them responded that they did not possess such information, and only 1% of them responded that student non-preparation is an obstacle to company development.

So even from this analysis we can conclude that student non-preparation is a major obstacle, as over 90% of them responded that partially and positively affecting student non-preparation affects the development of the company.

In the following we will present the analysis of whether the companies have data on labor market requirements in the region.

3.35. Market demands in the field of technology

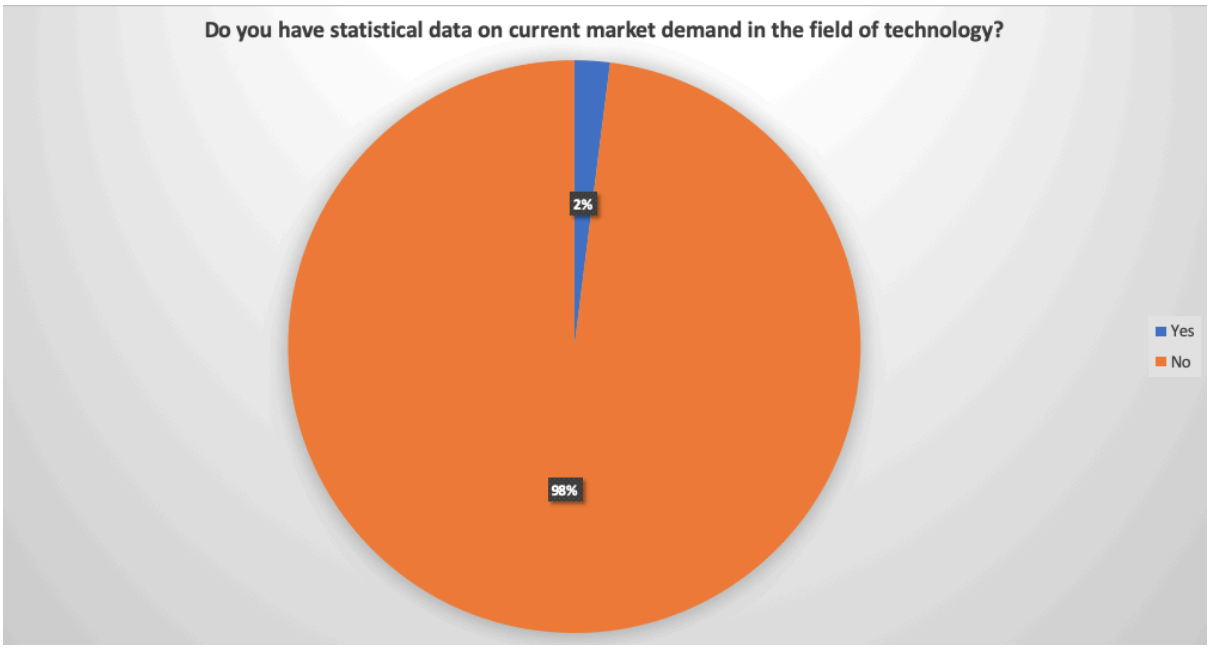


Figure 46. Market demands in the field of technology

In Figure 46 we present the results of whether companies have statistical data on labor market requirements in the field of technology. According to data obtained from companies and processed, 98% of them responded that they did not possess information on labor market demands in the field of technology, and only 2% of them responded that they possessed this information. Of course our research will be of great help in this regard as it will possess statistical data on labor market requirements in the field of technology.

We have finally presented the companies opinion on Automated model comparing between labor market demands and university curricula.

3.36. Companies on “Automated model comparing between labor market demands and university curricula”

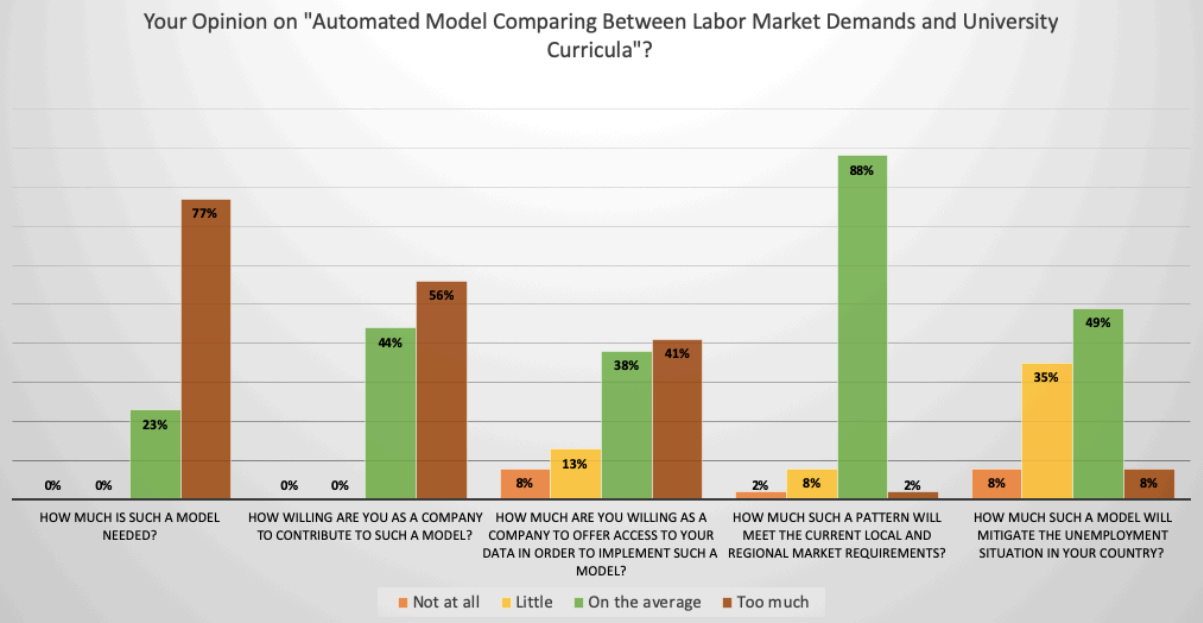


Figure 47. Companies on “Automated model comparing between labor market demands and university curricula”

Figure 47 presents an analysis of the companies' opinion about the automated model which will make a comparison between the labor market requirements and the university curricula. As part of our research it has been in our interest to know how much such a model is needed, how much companies will contribute to implementing such a model, how much companies will offer open access to their data, how much such a model will meet market demand, and how much such a model will improve unemployment in our country.

According to the processed results 77% of companies responded that such a model is very necessary, and 23% of them responded that on average the model is needed. Also, 56% of them responded that they would greatly contribute to the implementation of such a model, and 44% of them responded that on average they were able to contribute to the implementation of such a model. Of the companies that are able to offer open access to their data, 41% responded that they were very capable of delivering, 38% responded that they were able, on average, 13% responded that they would provide access small in data, and 8% of them responded that they would not provide access to their data at all.

Of such a model that would meet current market demands, 2% of them responded that it would meet many market demands, 88% of them responded that on average the model

would meet labor market requirements, 8% responded that they would meet very little, and 2% responded that such a model would not meet labor market requirements.

And finally, as such a model would improve the unemployment situation in our country, 8% of the companies responded that it would greatly improve, 49% of them responded that on average it would improve the unemployment situation in the country, 35 % of them responded that it would slightly improve, and 8% of them thought it would not improve the unemployment situation in the country at all.

Based on these data, according to the companies that responded to these questionnaires, it is a great motivation for us to continue our research in this direction, and to build a model that will be able to compare between requirements. of the labor market and curricula offered by the university.

3.37. Conclusion

So in the third chapter we have seen results that are derived from the survey of academic staff and students in public and private universities in Kosovo, and from private companies in the field of technology.

We started from the academic staff owned by public and private universities in Kosovo and we have seen a greater number of professors who possess the last two academic calls is needed.

Then we see the preparation of the students they possess when enrolling at the university, and the preparation with which they graduate from the university. Based on the results we have seen that the preparation is not the same, and we have made comparisons between the preparation of the students and the waiting time which the students have to become employed after completing the studies. Based on these analyzes we have seen that there is a strong link between these two variables, that universities should focus more on the professional preparation they offer to students.

Another point we have analyzed was also the part of the curricula offered by the universities, where we also presented the subjects in which the students encounter difficulties in the early stages of the studies. Then we have also seen how often these curricula are modified by universities, and we have also presented how many of the universities apply standard curriculum guidelines for the bachelor and master studies.

We have also seen that the number of agreements that universities have with private companies is moderately high, where part of these agreements have been the participation of business representatives in curriculum design offered by universities.

Also, in the third chapter, we have presented the amount of net income that student earns after graduating. And we have seen that the amount of net income is also on an average level, as advanced knowledge in: programming, databases, nets, etc. make the income in this field to be average.

How much is the same or similar the job that the students make compared to the field in which they have also studied, which we have shown graphically, where we have seen that only a small percentage of students do the same job or close to the field in which they have completed their studies. One of the main factors influencing not doing the same job in the field of studies is the lack of compliance between labor market demands and curricula offered by universities.

Finally, we presented how universities are ready to provide support for the implementation of such a model that will make a comparison between labor market demands and curricula offered by universities. Based on the results obtained, we have seen that a high percentage provide support in this regard by declaring that it is very important to create such a model. The support provided by private and public universities in this regard is related to the provision of open access to syllabuses which universities offer and the application of such a model to their system.

Also during this chapter we have seen how companies are satisfied with the skills students possess after completing university. We have seen what are the areas that there are vacancies offered by these companies, and we have seen what are the areas that need to be improved by graduate students.

During this analysis we have seen whether companies are familiar with the demands of the labor market, and in the end we have taken their opinion on the automated model that will make a comparison between the demands of the labor market and the curricula offered from university.

Therefore, we can conclude that the creation of such a model will have a positive impact on the state of the universities in Kosovo and in meeting the demands of the labor market in the field of technology.

It is important that creating such a model is not limited, so it can be applied to other countries in the region in order to find the level of adjustment between labor market demands and curricula offered by universities.

PART 4

4. Application of the model that will make comparisons between market demands and university curricula.

In order to achieve the implantation of our model we have identified the tools that will be used. Initially, it is the identification of sources of information that provide information on job offers offered in the field of technology. The initial idea was the information that would be used to deal with the bids of both our country and the region, but based on the small number of information provided by the websites in our country, we were forced to provide this information we get from European websites as the information capacity has been bigger.

Based on the information provided in our site from various websites, we have managed to get information on the bidding titles, which does not result in a proper analysis. In order for us to have a more accurate analysis, we will need to use websites that provide a large number of job offers in the field of technology. The information that comes into our hands is also the job descriptions that are provided by the bidders. Once the information source is identified, we start applying Web Scraping Techniques that will help us to extract the information in an automatic way. Since the preservation of the data will be completely unstructured, once the information is extracted, we will use text preview techniques, which will be used to avoid spaces and special characters that do not enter the work during our analysis . Once the text parsing is over, the wormhole corpus is considered ready for the application of machine learning techniques.

The same procedure will apply to the websites of universities that have published their syllabuses, where by means of web scraping techniques we will be able to extract the twin of the information that is included in the syllabi of the technology programs. After both the information corps are available, the same techniques will apply to the syllabus section in order to compare the text of the information on job offers, syllabi and technology programs offered by universities of the region in Kosovo, Macedonia and Albania. The language used to apply all these techniques is the Python language, as a powerful language and in the case of machine learning. In the following, we will talk about the sources of information that have been used in the recent years for the techniques that have been applied to achieve twin implementation of our model.

4.1. Source of information with job offers

As mentioned above, the target of the job offers information was the websites of our country and region. But, based on the small capacity of the information we have posted on our sites in the country, we have been forced to use European websites as they provide more information. In the following we will present the case when extracting information between the bids of the bids and comparing them with the bidding headline as well as extracting the description on the bundle of bids.

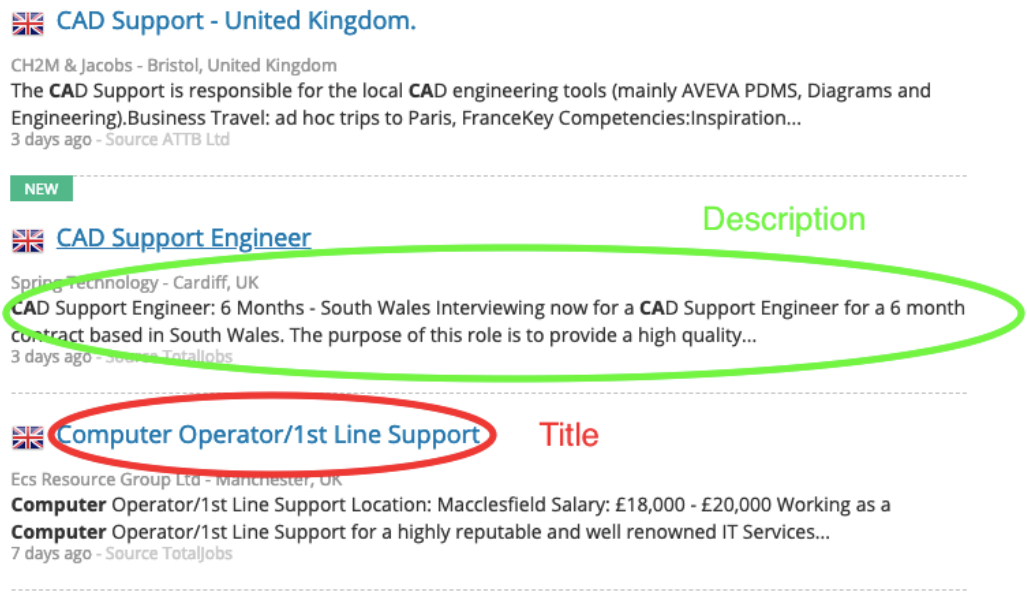


Figure 48. Unusable information

Just as it can be seen in Figure 48, this type of website is unsuitable for our case, as it does not provide us with information that enters our work. It can be seen that we can get information about the bidding title, as far as the bid description is concerned, there are no information that enters into the work. In this case, there are cases when websites do not allow us to apply web scraping techniques and we cannot get the information from that web site. Next, we will present a case where no web-scraping techniques are applied to a web site.

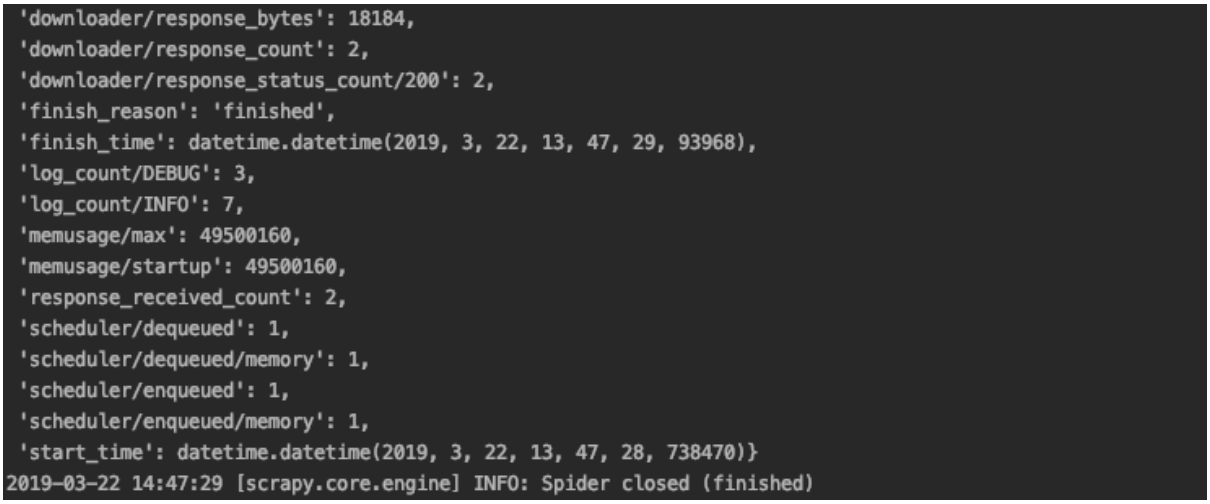


Figure 49. Unsuccessful scraping process

In Figure 49 it can be said that in a specific web page it is impossible to apply web scraping techniques due to the form of web site design. Based on the facts when some of the websites provide us with some information, and some of them with enough information do not allow us to apply web scraping techniques, we have been forced to look into the European website as well.

What is relevant to identifying information that will be extracted through web scraping techniques is that the code content that was used to construct the web site should be analyzed. To be more accurately explained, it is necessary to analyze whether the same template was used for all the links that are on this website, since the system will automatically visit all the links in order to extract the necessary information. At the beginning, the algorithm that is designed to extract the information is based on the tags defined in the web content. So if we have differences in the template model that is applied to the links within the webpage, then we will not be able to extract all the information from the webpage automatically. In the following, we will present the case when a website has enough information to enter our workforce.

Figure 50. Useful information from website

In Figure 50 it can be said that the specific website possesses enough information to implement our model. It is precisely the information of the title of the bid, the company which offers, the description of the applicant's knowledge of those who enter into the work. It is also important to note that this information web site also has its own customizable template, since the same template is used for all links within it. This construction of this web site enables and facilitates the release of automated information.

4.1. Scraping process

Next, we will present a sketch of how web scraping techniques will be applied to specific web sites. Where initially the initial link will be identified, then the identification of the tags on which the website was created will be identified. Once the tags are identified on the primary page, tags should be identified on the secondary page by identifying the bid title, the company that publishes the offer, and the description of the information that is required by that offer. In addition to each tag, you should also identify the links that contain the number of each website, since the website has a considerable amount of information.

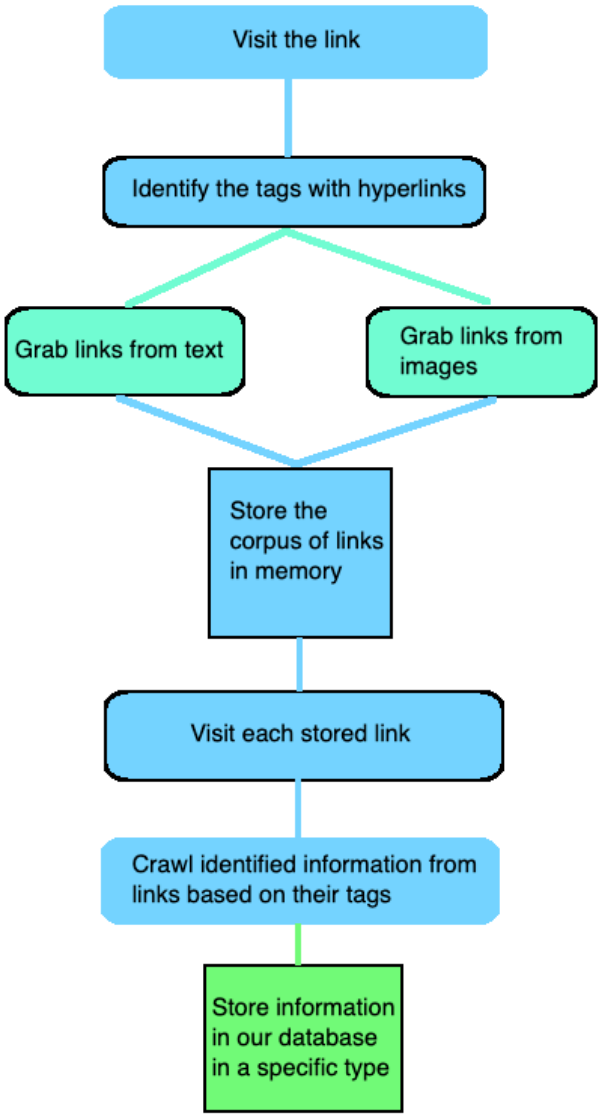


Figure 51. Scraping sketch

As we can see in Figure 51, the process of extracting information from the website goes through several steps. Initially, it is identified the primary link that will be visited by the system in an automatic way. Once the link has been successfully launched, it is the identification of the text and the photos that contain the hyperlinks to which we are retrieved within the site where we can get information on job offers. Once the links are extracted from the text and the photographer, they will be stored in a wormhole so that they can be visited later. Once they have been stored in large numbers, a link is clicked on, whereby tagging identifies specific textual content from the content of that web site. The content that is extracted from the website will be stored in the database as a specific information that will be used later on for

our automated model. The ability to save information is diverse, but the Json format is a format that is more adapted to our latest approach to applying machine learning techniques. In the following, we will present examples from identifying web pages, continuing with the algorithm that was used to automatically extract the information.

Figure 52. Identified tags

In Figure 52 you can indicate that the class is identifying, so that the link is extracted from it. This link is stored in the inside of the text and inside the image and we need to retrieve, store it in memory so that we can later visit this link. Once the link is saved and clicked, the system will be redirected to the website, which includes a lot of bullet information. Below we will present the case for identifying other tags that are within the site.

Figure 53. Identified elements

In Figure 53 you can see how the identification of the elements used to define the title, publisher, and job descriptions is used. Based on the web design template, within the "job-header" tag the element that was used to define the job offer title is tagged "h1", while the tag used to define the bidding publisher is tagged "h2 ". While the description of the work was used tag "h3" as well as other elements such as "p" and "ul". Once you have identified

each of these tags, you should also identify the last element of the links within which other websites are located. This part is known as "pagination" and allows the system to access the latest web page automatically, and at the moment it reaches the last page when the system is stopped ensuring that we do not visit the same site more than one .

Once all the elements are identified, we are ready to begin constructing the algorithm in the python language in order to execute and extract the information.

4.2. Algorithm for scraping process

As we mentioned earlier, the language that was used to execute the scraping process is Python. In the following we will present the algorithm divided into several parts in order to give explanations for the swirling of the part and finally to reach the test phase.

Figure 54. Import library

The initial part is the part of importing a scrapbook that is installed through the "pip" package, which installs and manages packages and applications written through the "python" language.

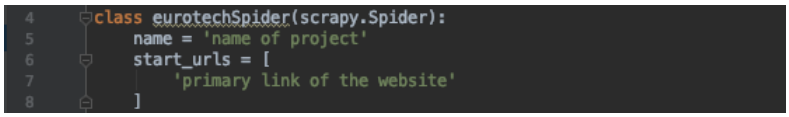


Figure 55. Class definition

As we can see in Figure 55, the second part of the algorithm is the part of the class definition which in our case is referred to as "eurotechSpider", continuing within the code with the name of the project which will later calls to execute via scrapy. Also in this part of the code may be the primary link used to visit the website in order to continue with the release of information.

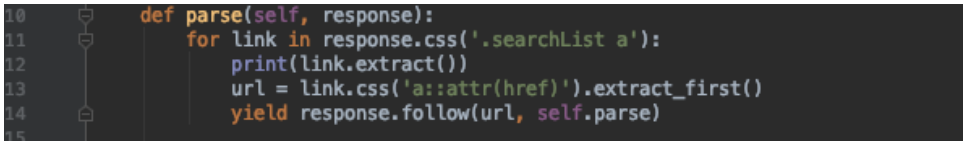


Figure 56. Parse definition

In Figure 56 it can be said that the definition of "parse" is going to be the output of the links that are defined. Initially, the "searchList" element is searched for all the "a" elements that result in links. Once you have identified all the links that are inside the webpage, they will be stored in a wormhole and will be visited each link that you can extract from the text or from the photo. There may be cases where the bidding of the bundle of bribes can save the link, but there are times when links can be found within the photograph that identifies the bidders.

```
15
16          vacancy = response.css('.job-header')
17          if vacancy is not None and len(vacancy) != 0:
18              description = '\n'.join(vacancy.css('li::text').extract())
19
```

Figure 57. Information that will be extracted

In Figure 57 you can see how it is defined as "vacancy" which will be checked within the "job-header" element, where it is the content of the description of the job offer. If the aforementioned element is not "none", then all parts of the website that contain the element "li" will be taken and the text will be included. In the present case, the text and the unlisted lists are used to describe the requirements of the published bidding. The last part of the code is the part where information is provided that can be extracted from the website and stored in a specific format.

The information we could get from the website was different, but in the case we needed only information that has the bulk of the job offer. Other information has been overlooked and has not been used in a way that does not adversely affect the system's performance during the execution of the algorithm.

```
20          yield {
21              'title': vacancy.css('h1::text').extract_first(),
22              'Company / Place': vacancy.css('h2::text').extract(),
23              'description': description,
24          }
25
26
```

Figure 58. Other types of information that will be extracted

At the bottom of the code was ordered the system for displaying the data we need. In this case we have identified that the bidding title was defined with the element "h1" and in what form will be extracted the text that covers this element. After extracting information from

element "h1", it is now extracting information from element "h2" which in this case holds the quotation to the bulletin bidder and publisher site. And the last part that is considered as the most important part is the description of the bulleted job offer, and now it will be invoked to be printed. All of the information that will be extracted will be stored in the "json line" format as a format that is appropriate for applying machine learning tweaks. Below we will present the algorithm in twin order to execute its execution and extract the appropriate information.

```
import scrapy


class eurotechSpider(scrapy.Spider):
    name = 'name of project'
    start_urls = [
        'primary link of the website'
    ]

    def parse(self, response):
        for link in response.css('.searchList a'):
            print(link.extract())
            url = link.css('a::attr(href)').extract_first()
            yield response.follow(url, self.parse)

        vacancy = response.css('.job-header')
        if vacancy is not None and len(vacancy) != 0:
            description = '\n'.join(vacancy.css('li::text').extract())

            yield {
                'title': vacancy.css('h1::text').extract_first(),
                'Company / Place': vacancy.css('h2::text').extract(),
                'description': description,
            }
```

Figure 59. Algorithm for scraping process

In Figure 59 is presented the algorithm that will be used to extract information from the specific web site. Since the algorithm is ready, its execution can be executed via the command "scrapy crawl name of the project". The change that needs to be made is the difference with regard to the format to which the data will be stored.

```
BOT_NAME = 'crawl'

SPIDER_MODULES = ['crawl.spiders']
NEWSPIDER_MODULE = 'crawl.spiders'

FEED_FORMAT = 'jsonlines'
FEED_URI = 'result.json'
FEED_ENCODING = 'utf-8'
```

Figure 60. The format of extracted information

In Figure 60, we present the change that is in order for our records to be stored in "jsonlines" format. Also part of this file is the file name that will be stored after executing the code in the

python language, as well as encoding as "utf-8" as an encoding commonly used in this format. Once all these changes have been completed, the execution of the algorithm will be executed and after the execution of the algorithm, the files will be kept in an unorganized form.

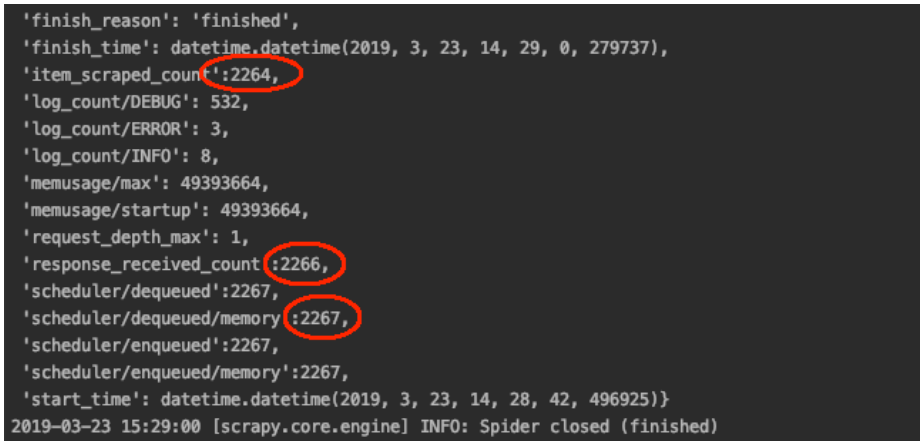


Figure 61. Successful scraping process

The successful completion of the process is shown in Figure 61, where we can say that after the execution of the code is the web page submission of over 2000 job offers. These data have been extracted automatically and stored in our system. After this process, the grants are now ready to undergo the process of getting rid of space and special characters in order to get the most out of the machine learning techniques.

4.3. Text processing

Since the downloads have been downloaded from specific websites, the same procedure will apply to the websites of universities that have published their syllabus programs in the field of technology. Knowing the nature of how web scraping works, there is a need to clean up the text that is being collected. The reason why it should be processed is that the files are collected from web pages, and they are also downloaded from other websites that are known as special characters. Next, we will show that our web pages still do not appear to be applied to text parsing.

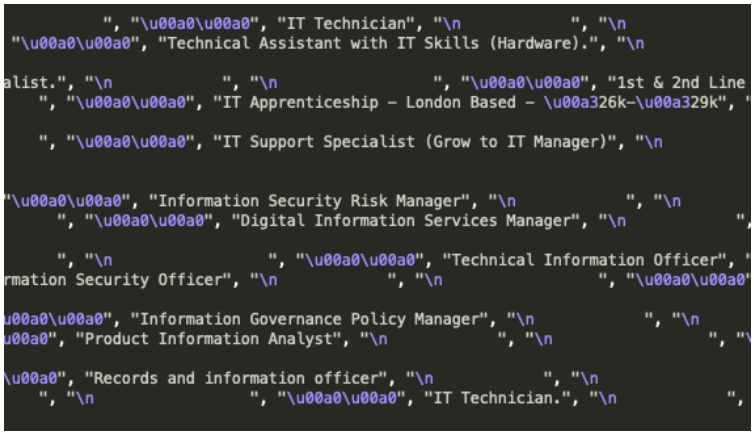


Figure 62. Unprocessed text

As can be seen in Figure 62, the files that are downloaded through web scraping are in an unsuitable format. After applying the "strip ()" and "replace ()" functions, there is the avoidance of spaces between words and rows, and also the download of special characters that are downloaded from the web site. In the following we will present the process of avoiding spaces and special characters from our document.

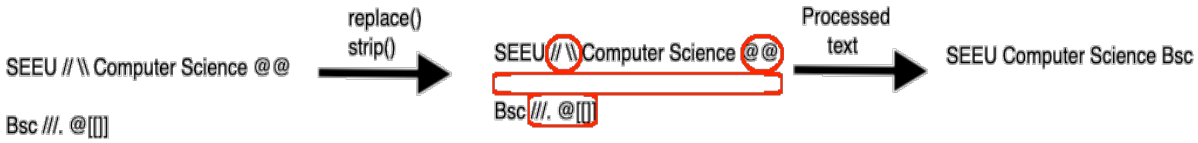


Figure 63. Text processing sketch

After applying the text cleansing techniques as shown in Figure 63, the text is considered ready for future techniques.

4.4. TF-IDF

The technique used to turn the word into a vector and to count each of them as soon as it is written in the report of all the words in the corpus is known as TF-IDF. This technique is a highly used and powerful technique that ensures that all of our corpus texts will be returned to a format that can be used for machine learning techniques.

There are several techniques that enable us to convert vector vectors, such as word2vec, term-frequency, etc., and all of them assure us that the text that converts to the vector is ready to apply the machine learning techniques .

As mentioned earlier, in terms of the main purpose of vector conversion, the TF-IDF has meant that each word is used as a whole, dividing it into the total number of words, and in

this form it extracts a statistic of each word. Next, we will present the formula for calculating TF-IDF according to mathematical calculations.

$$\mathrm{TF} = \frac{\text{COUNT WORD (DOCUMENT)}}{\text{TOTAL NUMBER OF WORDS IN DOCUMENT}}$$

So even as it is shown in formula, TF-IDF represents the number of each word divided by the total number of words that the corpus contains. Hence, the value of which we ultimately earn the value of each word, the greater the value that is earned, the greater the volume of that word used in that document will be the greater. However, we may already have cases where a word that is less relevant appears more often, and it cannot be considered that the word is more relevant. The concrete case is with the words "the", "this", "a", "an", "on", "in", known as stop words, and the ones used in more than one document cannot conclude what they are more relevant than other words. Therefore, in order to regulate such issues, it is required to be normalized, and this normalization is divided by dividing the number of those words into the total number of words in the document. In this way, the value of the words presented in our corpus will be reduced.

The function that is needed to divide between the number of words with the total of words that is inherent to the corpus also applies to the logarithmic function which calculates the logarithm of the values obtained from the total calculation. The formula that caused normalization of our records is as follows.

$$\mathrm{IDF} = 1 + \mathrm{LOG}\ \frac{\text{COUNT WORD (DOCUMENT)}}{\text{TOTAL NUMBER OF WORDS IN DOCUMENT}}$$

In the above formulas we can say that normalization is achieved by calculating the logarithm of the gained values and adding value 1 to avoid specific cases that do not reach negative values due to the number of words we have in the corpus. But in this case, the value of the above mentioned stop words has been normalized.

After computing with this formula we will get values as vectors that represent the input of each word based on the score of the tf-idf that is finally reached. The bigger the value of tf-idf the greater the wavelength of the words and the smaller the value of the tf-idf the smaller the waveguide of that word.

In both of the above formulas we have computed TF (Term Frequency) and IDF (Inverse Document Frequency) calculations, and to find the final tf-idf estimation at any time we can calculate the output between TF and IDF.

$$\mathrm{TF} - \mathrm{IDF} = \mathrm{TF}(\mathrm{SCORE}) * IDF\ (SCORE)$$

Below we will present an illustration of a simple example that implements the operation of the other calculations that assure us that we obtain accurate TF-IDF values of a document. We will use a corpus with a small number of words in order to see which words will be given greater power by the system.

In order to test the algorithm of the tf-idf we will use a corpus of 5 words where most of them have to do with one another. The terms used for illustration are:

- ***"This testing process is expensive"***
- ***"Testing an algorithm gives huge support"***
- ***"This software is very complicated"***
- ***"The learning process is becoming more complex"***
- ***"This software is very secured"***

Based on frequently used words, in this case, the bigger the word would be "this", "the", "is", "a" because they are written more. But based on the algorithm offered by tf-idf, this word should be smaller because, as mentioned above, these are known as stop words. Below we will present the result that was derived after the calculation of tf-idf on our list of keywords.

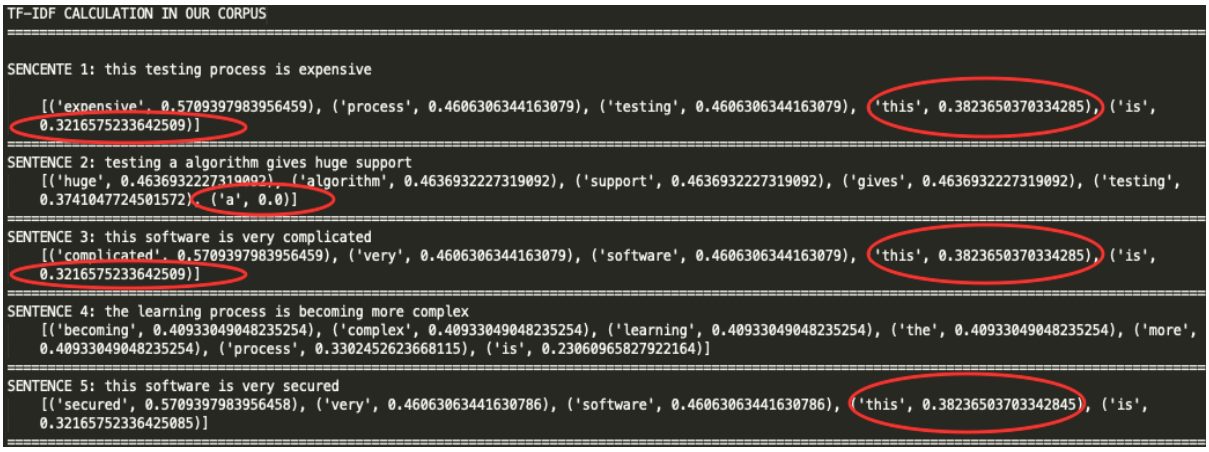


Figure 64. Tf-idf calculation

In Figure 64, we can show the results that were generated after the execution of the algorithm which made the calculation of the tf-idf of the words we set. Most widely used words are the "this", "the", "a" stop words, but even the most important word has been added to other words. In the first case, the words "expensive", "process" and "testing" were given greater weight than the words "this" and "is" which have a weight of 0.38 and 0.32.

Also in the second sentence we have the word "algorithm", "support" and "a", and as can be seen in Figure 43, the words "algorithm" and "support" were given the greatest significance of "a" it has a weight of 0. So in all cases, stop words are given less than the other words. This will give you even bigger results when you have a bigger body of words. Below we will present a sketch of how the algorithm functions which will compute the tf-idf.

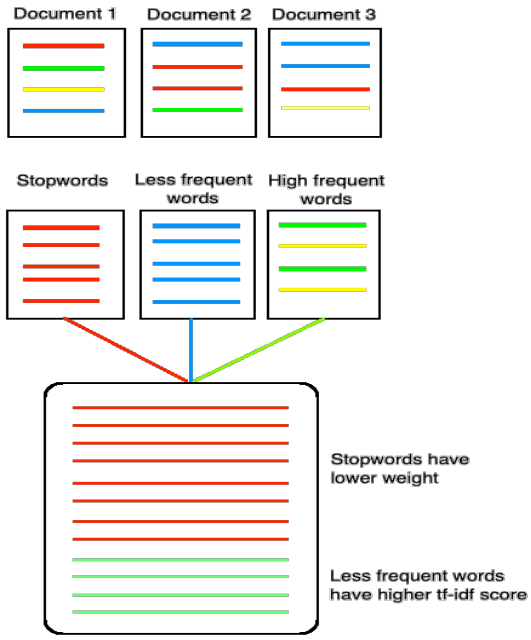


Figure 65. Tf-idf sketch

Just as it is shown in Figure 65 in the sketch, depending on the number of words that depend on their size. We have 3 documents that include stop words, words that have a large frequency and those with lower frequencies. And as you can see from all the documents, a document is created that contains all the words. The words that are placed in this document are classified into words that have the smallest and most widespread phrases.

4.5. Jaccard similarity

Twinning which are converted into numeric data and in the vector are ready to be applied to the comparison procedures since these will be the supply of job offers and study programs provided by universities in the field of technology.

Jaccard similarity is one of the techniques used for comparing two different datasets. Measuring the level of the similarity between two different datasets is calculated by comparing the two data sets. The level of similarity between the datasets by Jaccard similarity is 0 if there is no similarity, up to 100 if the similarities are identical. So the greater the Jaccard similarity coefficient, the greater the similarity between the two datasets. Next we will present the formula that computed the Jaccard similarity calculation to continue with its commentary.

$$\text{Jaccard}(\text{D1}, \text{D2}) = \frac{\text{D1} \cap \text{D2}}{\text{D1} \cup \text{D2}}$$

So, as can be seen in the above formula, calculating the Jaccard coefficient is calculated by dividing the dividing line between Document 1 and Document 2 and union of document 1 and document 2. If you want to sum up more than one of the above formulas Jaccard the coefficient we will have.

$$\frac{\text{D1} \cap \text{D2}}{|\text{D1}| + |\text{D2}| - \text{D1} \cap \text{D2}}$$

So we have a formula that represents the division between the intersection of Document 1 and Document 2, and the subtraction between Document 1 and Document 2 and the intersection between Document 1 and Document 2. The following figure will show graphically how Jaccard similarity works.

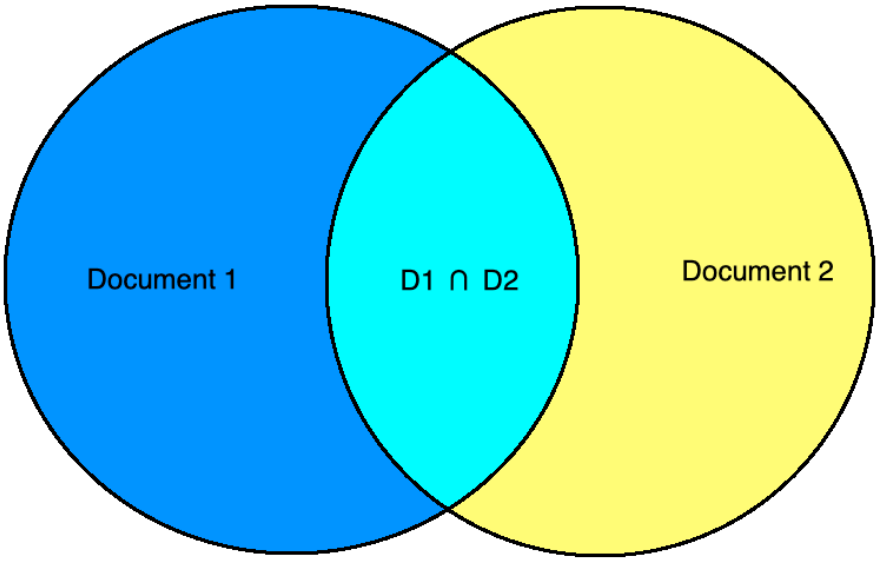


Figure 66. Intersection of documents

In Figure 66 we have graphically presented the intersecting case between the two documents, when the first document word is also in the second document. What value will be saved and used to divide the union of the two documents.

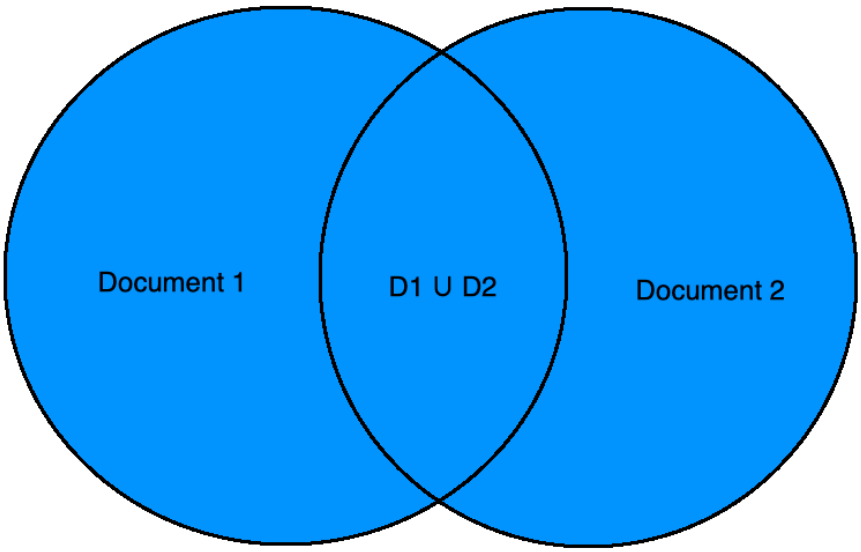


Figure 67. Union of documents

Figure 67 is a graphical representation of the two documents that presents the association of all the words found in the first document and in the second document. Also, these values will be saved as they will be divided by the value of intersection of Document 1 and Document 2. As Jaccard's similarity was explained, the following is illustrated by an example of how Jaccard similarity works.
To illustrate Jaccard similarity we will use two verses resulting from two different documents.

- ***Document 1: {1,3,5,6,7}.***
- ***Document 2: {0,2,4,5,6,7,8,9}.***

So, as we can see, there are two documents with different strings that after applying Jaccard similarity will equal the likeness of the same.
The first step is to calculate the intersection between Document 1 and Document 2, which will be the characters that are in Document 1 and Document 2.

$$(\text{Document 1} \cap \text{Document 2}) = (\{1,3,5,6,7\} \cap \{0,2,4,5,6,7,8,9\})$$

After intersection of Document 1 and Document 2 we will get the following value:

$$(\text{Document 1} \cap \text{Document 2}) = (5,6,7)$$

Based on the calculation of the intersection between Document 1 and Document 2, there are 4 characters that are in the first document but are also in the second document. Since we have the number of characters that are in the same two documents, then we will also calculate the union of Document 1 and Document 2.

$$(\text{Document 1} \cup \text{Document 2}) = (\{1,3,5,6,7\} \cup \{0,2,4,5,6,7,8,9\})$$

After union of Document 1 and Document 2 we will gain this value:

$$(\text{Document 1} \cup \text{Document 2}) = (0,1,2,3,4,5,6,7,8,9)$$

Based on the union calculation between Document 1 and Document 2, there are 14 characters that are in both documents but not duplicated two times the same character. Since we have both values, they are ready to apply to the formula with which Jaccard similarity is calculated.

$$\text{Jaccard}(D1, D2) = \frac{3}{|5| + |8| - 3} = 0{,}3 * 100 = 30\%$$

So from the calculation of Jaccard similarity between Document 1 and Document 2 with different words, it results that the text within each of these documents is the same with a value of 30%.

In some cases, depending on the calculation of the similarity between the two datasets, it is also necessary to calculate the distance between two datasets or documents which in their entirety include different words.

Below we will present the case of illustration of Jaccard similarity in python language after the transformation of the agitators.

```
1  from math import*
2
3  def Jaccard(d1,d2):
4
5   intersection_of_numbers = len(set.intersection(*[set(d1), set(d2)]))
6   union_of_numbers = len(set.union(*[set(d1), set(d2)]))
7   return intersection_of_numbers/float(union_of_numbers)
8
9  print Jaccard([1,3,5,6,7],[0,2,4,5,6,7,8,9]) * 100
```

Figure 68. Jaccard similarity algorithm

Figure 68 shows the algorithm that computed the Jaccard similarity calculation between two number ranges. Initially, the variables d1 and d2 are declared as two different documents, then the intersection_of_numbers are declared as intersection between the two documents. Once stated intersection_of_numbers, union_of_numbers are declared among the above mentioned documents. After calculating the intersection and union, calculating the Jaccard similarity function is the division between the two values. The value gained is 30% which can be referred to as the similarity between documents that cover the range of different numbers.

4.5.1. Jaccard distance

In order to calculate Jaccard distance, use the preliminary formula used to find Jaccard similarity. Jaccard distance as well-known as Jaccard dissimilarity is calculated by this formula:

$$\text{Jaccard distance}(\text{D1}, \text{D2}) = 1 - \text{J}(\text{D1}, \text{D2})$$

In the above formulas it can be said that the calculation of Jaccard distance is calculated by comparing the value of equation 1 and Jaccard similarity found to be greater. The formula will continue in this form:

$$\frac{|\text{D1} \cup \text{D2}| - |\text{D1} \cap \text{D2}|}{|\text{D1} \cup \text{D2}|}$$

So as we can see Jaccard distance is the division between the D1 and D2 union and the intersection of D1 and D2 and the union of D1 and D2. Below we will present the illustration of the same two previous documents.

$$\text{Jaccard distance}(\text{D1}, \text{D2}) = 1 - \frac{3}{|5| + |8| - 4} = 1 - 0{,}3$$

$$= 0{,}7 * 100 = 70\%$$

Based on the calculation that is based on Jaccard distance, it can be seen that Jaccard similarity between D1 and D2 is 30%, while the distance between both documents is 70%.

Next, we will present the Jaccard distance illustration even in python language after the agitation building.

```
#Jaccard Distance calculation
from math import*

def Jaccard_distance(d1,d2):

 intersection_of_numbers = len(set.intersection(*[set(d1), set(d2)]))
 union_of_numbers = len(set.union(*[set(d1), set(d2)]))
 return intersection_of_numbers/float(union_of_numbers)

print (1 - (Jaccard_distance([1,3,5,6,7],[0,2,4,5,6,7,8,9]))) * 100
```

Figure 69. Jaccard distance algorithm

Figure 69 shows an algorithm that computed the Jaccard distance calculation or Jaccard dissimilarity. As mentioned above, the formula that computed the distance between two documents with Jaccard distance is the same as Jaccard similarity, but is deducted by the value 1 with the result obtained by Jaccard similarity.

4.6. Cosine similarity

Since the model we will analyze real-time data from textual content documents, then the application of cosine similarity is required. Cosine similarity, from the contrast between Jaccard similarity which deals with the binary data of one set, it addresses the real data that can be textual.

Cosine similarity is one that is applied to measure the level of similarity between the two different documents.

The shape of Cosine similarity calculation is by converting the vowels into the vector, and by measuring the distance between two vector vectors that are projected into different axes in the X and Y axes.

Based on the shape of the cosine value, we have to know that two vector with the same orientation have cosine 1, and two vector with completely opposite directions have cosine -1. While the vector having 90 degrees with each other, one in the X axis and the other in the Y axis at the cosine is 0. Cosine similarity as a measurement method of similarities between the documents banged between positive values of 0 to 1. The values that the vector receive depend on the frequency as long as those words appear in the document, then the cosine similarity method allows us to find similarities between those vectors. The number of vectors that can be located in the multidimensional space is not limited, as does the number of comparisons between these vectors being limited.

4.6.1. Cosine similarity illustration

Below we will present some examples of how Cosine similarity works in order to calculate between two different documents. The Cosine Similarity Illustration will be used using several different documents that include different sentences in their interior. The documents to be used are:

- ***D1: Market demands are filled by university programs.***
- ***D2: Students are interested for our university programs.***
- ***D3: Market demands are too high.***
- ***D4: University programs should be very appropriate.***

We have four different documents that will be anonymous and compared with each other in order to calculate the similarity between them.

In order to continue with calculations of Cosine similarity, it is primarily the order of the words in all the words, ensuring that the words will not be written. The keywords that are used in the above mentioned documents are: *appropriate, are, be, by, demands, filled, for, high, interested, market, our, programs, should, students, too, university, very*. Each of the most listed keywords will be checked for how long each document has been submitted. Variously than each word has been presented, it will also be the value that it gains in placing it in the multidimensional space.

Once the table has been created and the numerical values are obtained for each word, then the mathematical calculations of cosine similarity will be made. The values that will be earned are presented as vector for each document. So, each document will be a vector, and once the results are obtained, the calculation will be similar to each of the documents. At the end of the calculation, you will also apply the tf-idf method to the first values as to how the results change because you are given a smaller value of the many words that are written. This will help us greatly because it can be said that in the documents that are the calculations there are words that are written and, as mentioned above, the fact that the words are written does not mean that the words are most relevant. Below we will present a table with words that are used in four documents.

Table 1. Words used for cosine similarity illustration

	D1	D2	D3	D4
appropriate	0	0	0	1
are	1	1	1	0
be	0	0	0	1
by	1	0	0	1
demands	1	0	1	0
filled	1	0	0	0
for	0	1	0	0
high	0	0	1	0
interested	0	1	0	0
market	1	0	1	0
our	0	1	0	0
programs	1	1	0	1
should	0	0	0	1
students	0	1	0	0
too	0	0	1	0
university	1	1	0	1
very	0	0	0	1

In the above table are presented all the words used in the above mentioned documents and the number of times that these words have been completed in all the documents. We have different occasions when one word is presented in one of four documents, and there are cases when a word is presented in three documents from four of them. Since we have numerical values, we will continue to parse vector values for each document, because as each document can have vector values. The resulting vector values will then be applied with the formula that computes the approximation of cosine similarity.

Table 2. Vector values of documents

	Vector value
D1	{0,1,0,1,1,1,0,0,0,1,0,1,0,0,0,1,0}
D2	{0,1,0,0,0,0,1,0,1,0,1,1,0,1,0,1,0}
D3	{0,1,0,0,1,0,0,1,0,1,0,0,0,0,1,0,0}
D4	{1,0,1,1,0,0,0,0,0,0,0,1,1,0,0,1,1}

The following table presents the value of each document that is presented as a vector based on the translation of the words in each document. Calculation of cosine similarity will be made:

$$\cos similarity = \frac{\mathrm{D} \cdot \mathrm{B}}{||\mathrm{D}||\,||\mathrm{B}||}$$

As can be seen in the above formula, cosine similarity represents the division between the two product variables, in this case Di and B, and the multiplicity of the swatch vector variables obtained for both documents. Respectively, the above formula will be blurred and after the breakthrough it will take this form.

$$\frac{\sum_{i=1}^{n} \mathrm{D}i\ \mathrm{B}i}{\sqrt{\sum_{i=1}^{n} \mathrm{Di2}}\ \sqrt{\sum_{i=1}^{n} \mathrm{Bi2}}}$$

Based on the above formulation which computed the values of the highest obtained vector, we will find similarities between the most recent documents we have presented. In the following we will present the calculation of the first and second documents, then not submit the calculation for all the documents, in the end will appear next to the table that contains all the results between each document.

$$\cos similarity = \frac{\mathrm{D1} \cdot \mathrm{D2}}{||\mathrm{D1}||\,||\mathrm{D2}||}$$

$$\frac{[0,1,0,1,1,1,0,0,0,1,0,1,0,0,0,1,0] * [0,1,0,0,0,0,1,0,1,0,1,1,0,1,0,1,0]\mathrm{T}}{Sqroot(1+1+1+1+1+1+1+1) * Sqroot(1+1+1+1+1+1+1+1)}$$

In the above calculation it can be said that the calculation of the dot product between the vector of the first and the second document vectors. This value is divided by the square root of each of these values. In the square root, the values are set to 1 because of the space in the document so that all the formulas are presented to us. After calculating these values, the values that we will earn are:

$$\frac{1*1+1*1+1*1}{\sqrt{7}*\sqrt{7}}$$

Also, more than just counts numbers greater than 0 because of the space you need for all the values. It is also the calculation of the values by which the value gained with the square root of the number 7 is multiplied by the same value. Since both of the two documents have a vector amount of 7 at both, their calculation is then calculated and then their output.
After the calculation, the values that have been earned are:

$$\frac{3}{2.6*2.6}=\frac{3}{6.67}=0.44$$

As well as the value obtained after calculating the cosine similarity between the first and the second document is 0.44, which can be calculated as a similarity between the two documents of 44%. Next, we will compute all the documents with each other to present one of the table's results in the form of one matrix.

Table 3. Cosine similarity calculation

	D1	D2	D3	D4
D1	1	0.44	0.51	0.59
D2	0.44	1	0.17	0.34
D3	0.51	0.17	1	0
D4	0.59	0.34	0	1

Based on the calculations that have been made for all the above mentioned documents, the values of cosine similarity for all the documents are presented. Just as it can be seen in the table above, the great similarity is between the first and the second document. This percentage reaches a value of 0.59 of cosine similarity, or else it may appear as 59% of the similarity between them. A small percentage of adjustment is between the first and the third document, where we have a similarity of 51%. While two documents that did not appear to be similar to each other, based on the results obtained are the third document and the second document since it has rated 0.

Below we will present the analysis of cosine similarity between these documents but with the application of the tf-idf in order to make the normalization of the data. This normalization we have mentioned is even higher when there are some words that have been written many times in the text. Therefore, you will be given the smallest amount of words you need when using the tf-idf methods.

Initially, the calculation of TF or term frequency is calculated by calculating how many times each word appears in the report with all the words in the document. The number of words that are enclosed within a document is then used to calculate its logarithm. The logarithm of these values is known as IDF, and once both values are found, then the product is calculated between them in the order of the tf-idf.

This value is considered as a normalized value and smaller compared to the direct values that were later calculated in the cosine similarity. The next step is to calculate the cosine similarity with the values that are normalized, and of course the final values that we will gain in our calculations will be different and smaller compared to the previous values. A large difference can be found in documents with a large number of spelled words.

Table 4. Tf calculation for each word

TF calculation				
	D1	D2	D3	D4
appropriate	0/7	0/7	0/5	1/6
are	1/7	1/7	1/5	0/6
be	0/7	0/7	0/5	1/6
by	1/7	0/7	0/5	1/6
demands	1/7	0/7	1/5	0/6
filled	1/7	0/7	0/5	0/6
for	0/7	1/7	0/5	0/6
high	0/7	0/7	1/5	0/6
interested	0/7	1/7	0/5	0/6
market	1/7	0/7	1/5	0/6
our	0/7	1/7	0/5	0/6
programs	1/7	1/7	0/5	1/6
should	0/7	0/7	0/5	1/6
students	0/7	1/7	0/5	0/6
too	0/7	0/7	1/5	0/6
university	1/7	1/7	0/5	1/6
very	0/7	0/7	0/5	1/6

The following table presents the TF calculation for each word, which represents the number of words that is written in relation to the total number of words in the document. In the first and second documents we have seven words, while in the third document we have six words and in the fourth document there are six words. The number of each word is divided by the total number of these words. Next, the calculation of IDF or inverse document frequency will be done, which in turn directly calculates the weight of the word swatch as much as it is relevant.

Calculating the inverse document frequency will be calculated by calculating the sum of all values for each word. In this case, the calculation of the values of the first document will be calculated.

Table 5. Calculation of log for words

	IDF calculation	Log	Log
appropriate	log(4/1)	log(4)	0.6
are	log(4/3)	log(1.3)	0.1
be	log(4/1)	log(4)	0.6
by	log(4/2)	log(2)	0.3
demands	log(4/2)	log(2)	0.3
filled	log(4/1)	log(4)	0.6
for	log(4/1)	log(4)	0.6
high	log(4/1)	log(4)	0.6
interested	log(4/1)	log(4)	0.6
market	log(4/2)	log(2)	0.3
our	log(4/1)	log(4)	0.6
programs	log(4/3)	log(1.3)	0.1
should	log(4/1)	log(4)	0.6
students	log(4/1)	log(4)	0.6
too	log(4/1)	log(4)	0.6
university	log(4/3)	log(1.3)	0.1
very	log(4/1)	log(4)	0.6

In the next table can be used to calculate all the values that are now readily computed with the values earned by TF. The formula to be used for the calculation of TF-IDF is the above-mentioned formula.

$$\mathrm{TF} - \mathrm{IDF} = \mathrm{TF}(\mathrm{SCORE}) * IDF\ (SCORE)$$

Based on this calculation for each word we have earned the TF score, it will be multiplied by the IDF score which was won by the IDF calculation. In this case, for the first "proper" keyword that has TF value in the first document 0/7, and IDF is 0.6, results with 0 since the output between the two values gives us the value 0. All other documents the values are 0 as with the calculation without TF-IDF, while the fourth document result is 0.1, since we have the

calculation between 1/6 * 0.6. And when comparing the value obtained without TF-IDF, which was 1, in this case the value is 0.1. Below we will present a table which presents all the calculations for the words from the previous documents.

Table 6. TF-IDF calculation

TF – IDF calculation				
	D1	D2	D3	D4
appropriate	0	0	0	0.1
are	0.014	0.014	0.02	0
be	0	0	0	0.1
by	0.04	0	0	0.05
demands	0.04	0	0.06	0
filled	0.08	0	0	0
for	0	0.08	0	0
high	0	0	0.12	0
interested	0	0.08	0	0
market	0.04	0	0.06	0
our	0	0.08	0	0
programs	0.01	0.01	0	0.01
should	0	0	0	0.1
students	0	0.08	0	0
too	0	0	0.1	0
university	0.01	0.01	0	0.01
very	0	0	0	0.1

For each word in each document we have the values that are normalized and have weighed less than the first case when the TF-IDF was not applied. The values we have achieved are positive values from 0 to 0.1, compared to the first case when the values were 1 and 2. As we mentioned above, these changes are even more when we have a number of twelve great for words that are inside a document. Since we have gained normalized values, we will now calculate the cosine similarity with the values so that the similarity between the documents is found.

$$\frac{[0.014,0.01,0.01] * [0.014,0.01,0.01]\mathrm{T}}{Sqroot(0.014 + 0.2 + 0.02) * Sqroot(0.014 + 0.32 + 0.02)}$$

After calculating the values then we get:

$$\frac{0.014 * 0.014 + 0.01 * 0.01 + 0.01 * 0.01}{\sqrt{0.23} * \sqrt{0.35}}$$

After calculating the values then we get:

$$\frac{0.000396}{0.27} = 0.014 * 100 = 0.14$$

Table 7. Cosine similarity with normalized values

	D1	D2	D3	D4
D1	1	0.14	0.18	0.7
D2	0.14	1	0.17	0.34
D3	0.18	0.17	1	0
D4	0.7	0.34	0	1

In the next table you can see the final results after the normalization of the data by TF-IDF techniques. As you can see, the bigger value that was 59% in the calculation after TF-IDF, after normalization this value has rush to 0.7. This has happened because the words that have a great deal of significance have been lowered by their weight being accounted for as less significant. We can also see that the documents that did not have the same D3 and D4 resemblance also resulted in 0%. While fewer documents are the D1 and D3 documents with 0.18, as well as D2 and D3 with 0.17.

Below we will present the graphic case of the words in the multidimensional step, where we have the identical identities, the smallest similarity, and the likeness of 0 among the words.

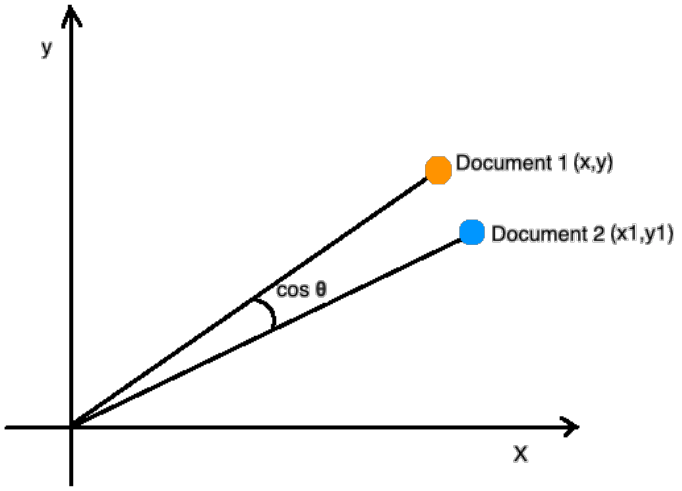


Figure 70. Identic documents

In Figure 70 we have presented the case when we have two documents that are located in the multidimensional space at the same point. As mentioned above, the position of the vector in the multidimensional space between the points x and y depends on the expression of the words which are in their entirety. In the case of this case, we have a large number of documents, since there is no question as to how much distance the documents are in their own, but the relevant ones are between them in the range of 0. We can easily conclude how many the smallest of the two documents, the bigger the similarity between them, always approaching the value 1.

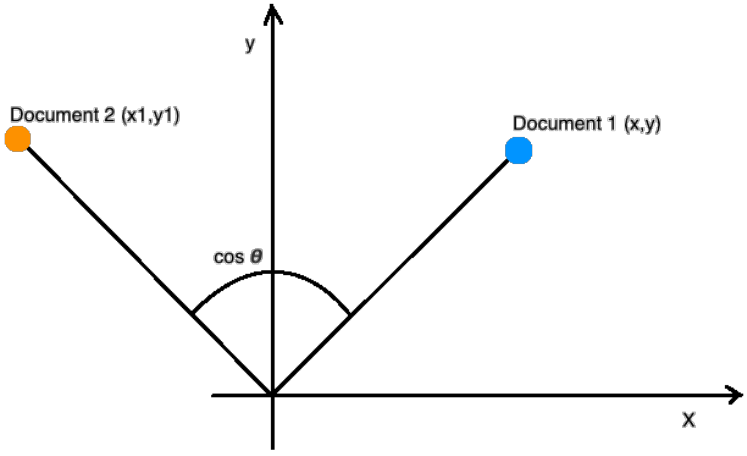


Figure 71. Different documents

In figure 71 there is a case when we have two documents that are not similar to each other, and the distance between them is bigger. As you can, the distance between document 1 and document 2 is approximately 90 degrees. And the bigger the distance between the two

documents is the closer we approach the value 0, which results in the smallest similarity between the documents.

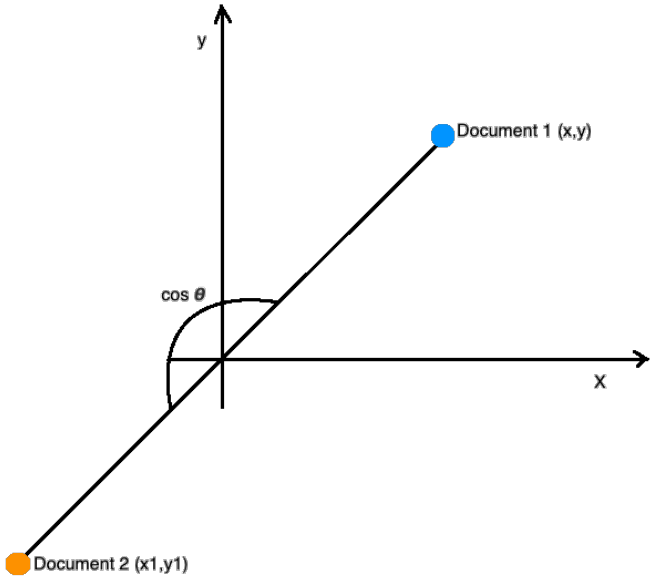


Figure 72. Completely opposite documents

In Figure 72 it is illustrated the case when we have two documents that are completely opposite to each other. As well as the first document, the first document is placed in the x and y axes at positive values, while the second document is placed at the negative values of the x and y constrains. Since the two documents are placed in diametrically opposite dimensions, the number of documents is larger, which can reach 180 degrees. With such a large angle, after calculating cosine its value will be negative. And when we have negative values that occur during the calculation of cosine similarity, then the similarity between them is calculated as the value of the value 0.

In the following we will present the case of automated model implementation which will be based on all the illustrations that have been up to now. As we have seen, the techniques were initially tested with little value to see if they work, so that they later come together in a single model and apply to larger documents. Once the automated modeling of the workforce and the curricula offered by universities is illustrated, we will present different analysis of the model by our model.

4.7. Application of the model.

Numerous analyzes have been used and are being proposed to address the requirements of the labor market with curricula offered by universities. Some analyzes that have been made so far are manual analyzes that are applicable to cases when we are faced with a large dynamics of data. Due to the timely and cost-effective factor, such analysis in any manual can result in even a small amount of exact results that can be derived from the analysis.

Therefore, the solution of this problematic as well as mentioned earlier is the application of an automated model which will be in a state of comparison between the labor market demands and the curricula offered by universities .

The model was based on mathematical calculations, which were applied through special algorithms through the python language. These calculations have to give us more accurate results in terms of tailor-made labor market demands with university curricula.

As mentioned above, the purpose of our model is to be able to provide analysis in other areas as well, and it can be applied in other areas as well.

In this chapter we will give examples and illustrations of the design of our model to different parts, ranging from the conversion of counts to numerical values and to the final analyzes that are derived from our model.

The characteristic of our model is that we can analyze two or more documents at the same time, and their length may vary. So the analysis that will be made of each document with each document. It is important to develop a corpus of data that will be used in order to normalize the data values that will be used for analysis.

This corpus was created by some of the words that were extracted from the downloaded text from websites that published information on job offers. Part of the analysis will be some of the universities in the region, where they will use the information that has been published by these universities regarding the curricula that they offer in the field of technology.

4.7.1. Sketch of our automated model

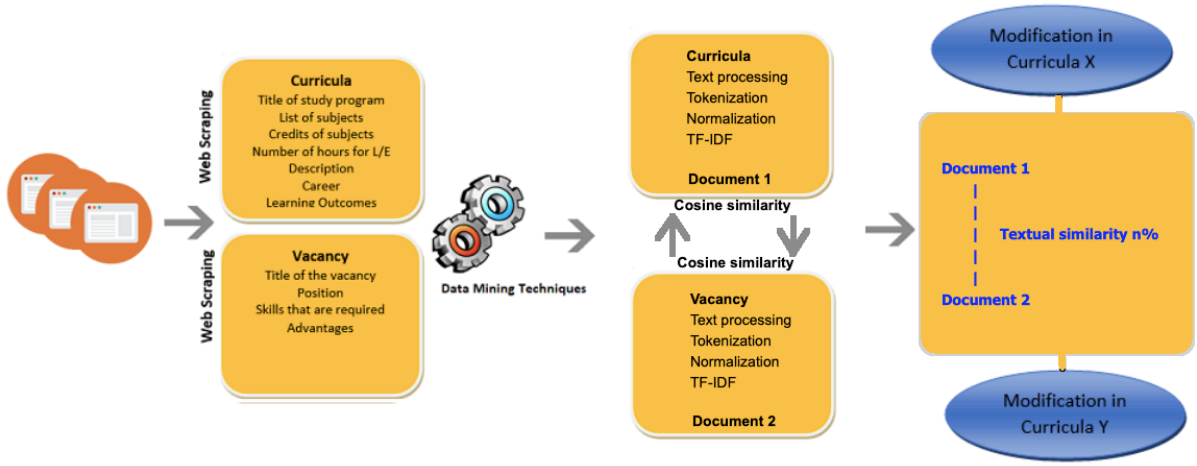


Figure 73. Automated model sketch

In Figure 73 we can see the sketch of our automated model that makes comparisons between market demands and university curricula. Compared with the model that was originally proposed as it can be, there is no change, except for the techniques that will be used for comparison. Which techniques that are used in our model are defined after research that is part of the field of machine learning.

The first step is the identification of websites that publish information on job offerings in the field of technology, as well as the identification of curricula of the study programs of the universities that will be part of the research.

Once the websites are identified, a web scraping algorithm will be applied that will visit the websites and extract the information that is specified.

The third step is processing the data, where it is cleared from duplicate spaces between characters and rows, as well as removing special characters.

The third step was to include the tokens of tokens, the normalization of the text, and the conversion of each word to vector values through TF-IDF.

The last step is the application of cosine similarity between two documents that are almost ready to be compared. The ultimate result is the textual similarity between the two documents based on the same words that are in both documents. Based on these results, we will be able to give recommendations about the program designed for proper implementation in order to fit the demands of the labor market.

4.7.2. Algorithm used to create the model

The algorithm used to create a model that will be divided into several parts between the requirements of the labor market and university curricula. The first part begins with the import of libraries that allow us to execute the algorithm that is constructed based on those libraries. In the following we will present each part of the algorithm that is used to comment on each part of it.

```
1  #Import of all libraries
2  from __future__ import division
3  from sklearn.feature_extraction.text import TfidfVectorizer
4  import math
5  import string
```

Figure 74. Import of libraries of automated model

In Figure 74 we have presented the code used for imported libraries that enable us to execute the functions used in our algorithm. As already mentioned in the code, at the beginning of the algorithm is the import of mathematical libraries, as well as the importation of TFIDF libraries from the 'sklearn' library. Which library will be used throughout the algorithm, since in our algorithm we have also created mathematical functions that are computed by division. Also in the bookstores that were imported came the 'string' booklet, which allows us to twin string functions.

As mentioned above, during the implementation of our model, we have been asked to use a corpus of data that will normalize the value of words that have been used most often in some cases. The words that are often used in our text, after application of the gift corpus will lower your weight compared to other words that are used less. Next we will present a part of the corpus which will later be used in the algorithm in order to make normalization.

This corpus comes with the words 'unigram' which are separated as single words that are extracted from the text extracted from websites that publish the bidding offers in the field of technology.

The list of information includes information on the words used, the number of frequencies used in the past, and their percentage in relation to all the words used in the text extracted from websites that publish job offers in the field of technology.

Word	Frequency	Percetage
technology	3599	9,127%
analyst	1832	4,646%
information	1605	4,070%
security	1497	3,796%
support	1146	2,906%
senior	988	2,506%
end	662	1,679%
user	633	1,605%
software	628	1,593%
head	620	1,572%
development	614	1,557%
technician	570	1,445%

Figure 75. Frequency and percentage of used words

Figure 75 is a list of words that were used throughout the text that was extracted from the web pages. As we can say, the word we use in comparison to all the words is the word 'technology', as compared to all other words, this word was used over 9%. Other words have a high level of use since most of them have a frequency of over 500 in the entire document. This corpus with what they would otherwise call 'unigram' will be incorporated within our algorithm in to be normalized with the words used in our document. Below we will present how we have managed to incorporate this word into our document.

```
ngram_freq = {
    'technology': 3599,
    'analyst': 1832,
    'information': 1605,
    'security': 1497,
    'support': 1146,
}
```

Figure 76. Ngram words

Figure 76 represents the filament corpus that was used in the case of normalization of the filaments. The corpus containing the list of words with the frequency of words in our case is declared as 'gram_freq'. In the code of the code, these words are used to divide each word into the number of words in the total. The main reason why we have decided to use a newcomer created by us because of the fact that the public offering provided by google, contain a large number of words that do not enter the work.

Once we have imported the library and created the corpus with our own data, now in our algorithm we will create the source of information which will be used for comparison. Based on the work we have done with our grants, we have available job offers and university

curriculum bidding documents. Also those documents were subjected to the processing of the text, where the application of special characters and special characters that did not enter the work were applied.

```
#Tokenization of data, and transformin in lowercase
tokenize = lambda document: document.lower().split(" ")

#Source of our data. Both of documents, vacancy and curricula
vacancy_document = open("/Users/Ylber/Desktop/eurotech/vacancy_document.json").read()
curricula_document = open("/Users/Ylber/Desktop/eurotech/curricula_document.json").read()

#The documents which are used in our anaysis
used_documents = [
    vacancy_document,
    curricula_document,
]
```

Figure 77. Documents that will be calculated

The procedures presented in Figure 77 are the disclosure of information sources and tokenization of documents, as well as the small-scale translation of all words. When we are reporting the source of the information, we have stated for both documents the destination where the documents which have been filed processing procedures have been stored. Also used is the 'used_documents' variable, which represents the list of all documents that will be analyzed by our model. As we have mentioned earlier, our model is not limited to the number of documents that will be used for comparison.

Tokenization of text is part of the three parts, the first part of the 'lambda', where we keep the document as an argument using the lambda function. Once we have saved the document through the lambda function, we will use this argument in the next step that is the conversion of all words into a lowercase. The process of converting the words of the corpus into the lowercase is used for all the words of the same weight, since they differ from the weight of words spoken in large letters and smaller ones. Once the transformation of all the words into a lowercase is done, they will split them by the split function.

After all these procedures, our documents are readily available to the TF-IDF technician, and converted into numeric or vector values. Below we will present cases of definition of TF, IDF functions, as well as other functions of normalization of the word weight of the corpus .

```
#Definition of TF based on library
def TF_of_documents(word, tf_document):
    return tf_document.count(word)
```

Figure 78. Definition of TF

As we have shown in Figure 78, we have illustrated even higher, so as we can say that in this part of the algorithm we have defined the term_frequency which will make the tokenization of documents, dividing each word in a row and by counted how many times each of them had been rewritten.

```
#The process of normalization
def first_normalization_technique(word, tf_document):
    count = tf_document.count(word)
    if count == 0:
        return 0
    return 1 + math.log(1 + count / (ngram_freq[word] if word in ngram_freq else 1) * 1000 )
```

Figure 79. Process of normalization

Normalization of data is very much needed since the files that are extracted from various websites also include words that are often used, and words spoken in English are known as 'stop words'.

In terms of what is said to be equated with other words, two normalization techniques apply. The technique used is the division of each word with the total of how much time that word was copied across the document. Once this calculation is made, this value is multiplied by the value of 1000 and the value is added 1. Once the data is normalized, the IDF or inverse document frequency will now be calculated.

```
#Definition of IDF based on library
def inverse_document_frequencies(tf_documents):
    idf_values = {}
    all_tokens_set = set([item for sublist in tf_documents for item in sublist])
    for tkn in all_tokens_set:
        contains_token = map(lambda doc: tkn in doc, tf_documents)
        idf_values[tkn] = 1 + math.log(len(tf_documents)/(sum(contains_token)))
    return idf_values

def tfidf(documents):
    tf_documents = [tokenize(d) for d in documents]
    idf = inverse_document_frequencies(tf_documents)
    tfidf_documents = []
    for document in tf_documents:
        doc_tfidf = []
        for word in idf.keys():
            tf = first_normalization_technique(word, document)
            doc_tfidf.append(tf * idf[word])
        tfidf_documents.append(doc_tfidf)
    return tfidf_documents
```

Figure 80. Definition of IDF

The IDF calculation is presented in Figure 80, where it can be used that words are used since they are normalized by prior technique. As we can say, in our algorithm we have stated 'lambda', which as an argument conserves the values that are calculated from preliminary techniques for normalization. Once saved, the IDF mathematical function is applied to those data. As already illustrated above, the function that computed the IDF is the division

algorithm between the word frequency and the total number of words in the document. Finally, to avoid as much as negative values, all 1 of the functions are added.

Once the IDF is calculated, we are now ready to calculate the TF-IDF as we have exemplified along the way. The TF-IDF calculation was calculated by computing the product between the TF score and the IDF score. The first technique of calculating the similarities between the documents is presented to this point, now we will present the declaration of functions that will make computations of cosine similarity.

Cosine similarity as a technique for comparison between documents, allows us to compare documents of different lengths, so we will present the steps that are used in order to apply cosine similarity techniques.

```
#Application of tfidfvectorizer
tfidf_from_sklearn = TfidfVectorizer(norm='l2',min_df=0, use_idf=True, smooth_idf=False, sublinear_tf=True, tokenizer=tokenize)
sklearn_score = tfidf_from_sklearn.fit_transform(used_documents)
```

Figure 81. Declaration of tfidf vectorizer

In Figure 81 you can use the second technique that made the conversion of documents that were initially declared in vector values. So the source of the information that will be used is stated as 'used_document', and are just some of the documents that will be converted to numeric values.

```
def cosine_similarity(vector1, vector2):
    product = sum(p*q for p,q in zip(vector1, vector2))
    extent = math.sqrt(sum([val**2 for val in vector1])) * math.sqrt(sum([val**2 for val in vector2]))
    if not extent:
        return 0
    return product/extent
```

Figure 82. Definition of cosine similarity

Since our data are converted to numeric or vector values, we can now apply the algorithm that will compute the cosine similarity calculation. Based on the illustrations that we have presented above, the calculation of cosine similarity is divided by dividing the multiplicity of vector with the square root of the plural of vectors. And since these calculations are made, this value will be kept as a value that compares the comparison between both documents.

By calculating that at this point we have the results of both techniques, then in the last part of the algorithm we will call the results obtained from both techniques in order to compare the results. Next we will present an algorithm that shows the results obtained by TF-IDF and cosine similarity.

```
#Comparison between cosine similarity and tfidf
results_tfidf = tfidf(used_documents)

compare_of_tfidf = []
for score_0, document_0 in number(results_tfidf):
    for score_1, document_1 in number(results_tfidf):
        compare_of_tfidf.append((cosine_similarity(document_0, document_1), score_0, score_1))

print(compare_of_tfidf)
```

Figure 83. Comparison between two methods

Figure 83 shows an algorithm that compares the results obtained from both of the above-mentioned comparison techniques. Initially, 'results_tfidf' was declared and then compared to the previous values. As we can also point to our algorithm, we've presented our results as document_0 and document_1 as well as score_0 and score_1. What we will finally get is the value of TF-IDF and cosine similarity between the two documents. Now we will present the results after executing the algorithm.

4.8. Commenting the results.

Following the application of the model we will be part of the analysis of the universities of the region which provide study programs in the field of technology. Also, as sources of information on job offers, selected web sites that publish competitions for the European market.

In the following we will present some cases of analysis in order to continue to comment on them. The analysis will be based on the job offers that have been published as well as on certain programs.

In the following we will present the analysis of each syllabus of the universities with job offers in order to present which syllabus has the most similarity with the market requirements. We will first present the graphs of the South East European University, and then proceed with the comparisons of the University of Prishtina and that of Tirana.

4.8.1. South East European University versus market demands

Figure 84 shows the comparison between South East European University syllabuses and labor market requirements.

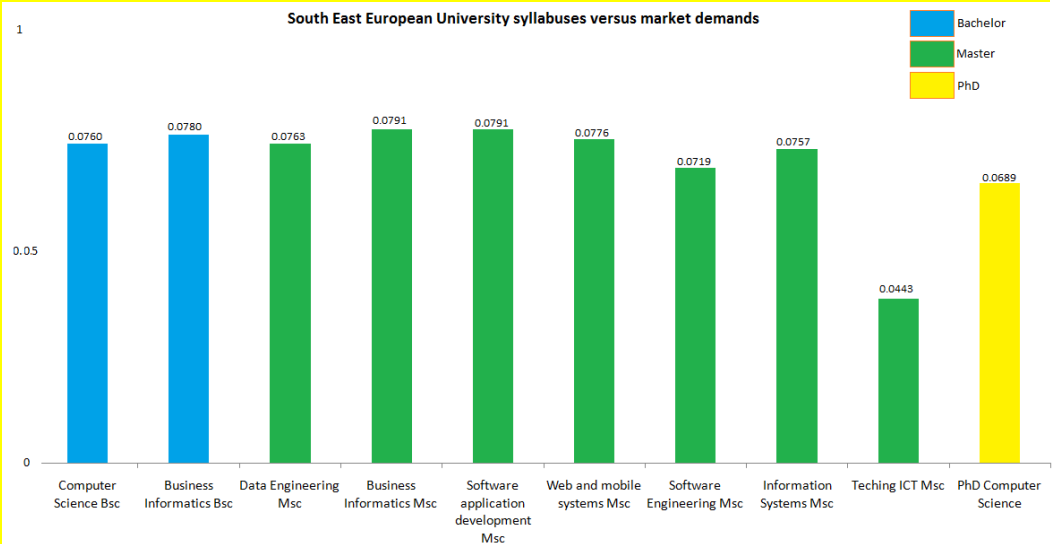


Figure 84. South East European University syllabuses versus market demands

As we can see in figure 84, our model compares European market requirements with syllabuses published on the South East European University website.

According to the analyzes that are derived from our model, we can see that there is a similarity to almost all syllabuses offered by this university. Starting with the Computer Science program we have a textual similarity of 0.0760 between the syllabus content and the labor market technology requirements. A similar adjustment is true for the Business Informatics program as we obtained a 0.0780 similarity between this program and the labor market requirements.

Such a comparison is made for other levels of study, master's and phd. As we can see in figure 84 for the Data Engineering program offered at the master's level we have a textural similarity between this program and the labor market requirements of 0.0763. One more similarity is for the other two programs offered at the master's level of Business Informatics and Software Application Development, where the similarity of the textual content between these two programs and the labor market requirements is 0.0791.

For the Web and mobile systems program we also have a rough textual similarity with other programs as the textural content of this program with labor market requirements is 0.0776.

Also for the Master Engineering Software program we have a similarity to other programs, where by calculating our model results a similarity of 0.0719. A great deal of similarity is also gained for the Information Systems program offered at the master level, where after calculating the similarity it turns out to be 0.0757.

Compared to the programs mentioned above, a smaller similarity has been gained between the textual content of the Master-level Teaching ICT program and the labor market requirements in the field of technology. After calculating the similarity using our model, this program turns out to have a similarity of 0.0443, which is a smaller similarity if compared to other programs offered by South East European University. Certainly during this chapter we will be able to identify what words are missing in these syllabuses by directly influencing the similarity to be smaller.

Studies that are also offered at South East European University are PhD studies. Of course, such an analysis has also been necessary for the program offered at this level in the field of technology. After calculations made using our model, the textual similarity between the PhD level computer science and the textual content of labor market requirements turns out to be 0.0689. Compared to similarities acquired for other levels of study, this program also results in a satisfactory similarity to labor market requirements. In the following we will present the analysis of the programs offered by the University of Prishtina for all levels.

4.8.2. University of Pristina syllabuses versus market demands

Since the purpose of our research has been to analyze the programs for the universities in the region, then such an analysis has certainly been done for the University of Prishtina by extracting information from its website. In the following we will present the graph with the analysis of programs and requirements of the labor market to continue with their commentary later.

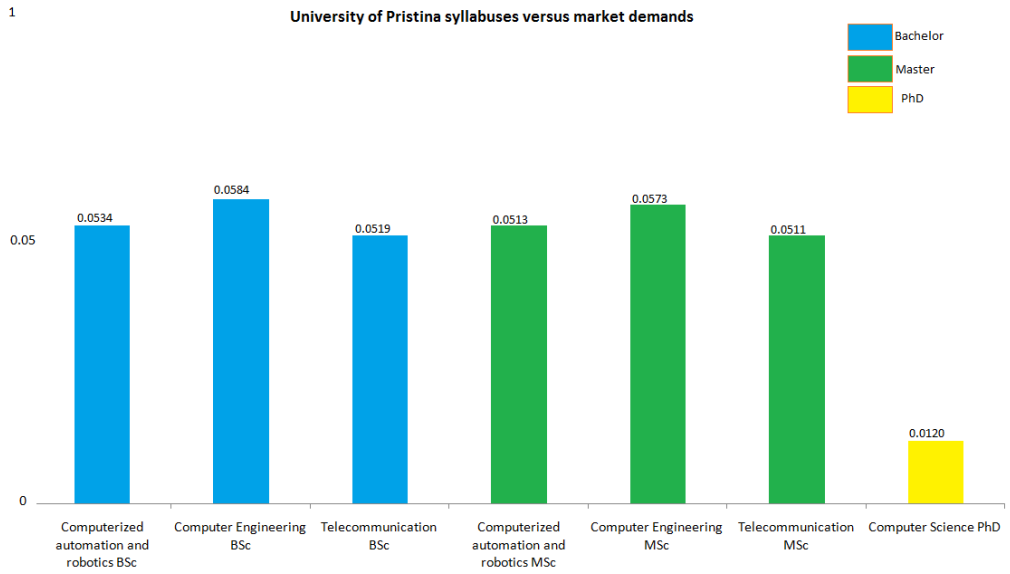


Figure 85. University of Pristina syllabuses versus market demands

The programs offered by the University of Pristina in the field of technology are Computerized Automation and Robotics, Computer Engineering and Telecommunication for bachelor degree. The same degree programs are also offered for the master's degree, while the Computer Science program is offered for the PhD level.

The information extracted from this University is derived through automated techniques, and used to analyze the similarity between them and the demands of the labor market.

For the bachelor's degree Computerized Automation and Robotics program we have gained a similarity of 0.0534. Also for the bachelor of Computer Engineering program we have acquired a textural similarity between this syllabus and the textual content of the labor market requirements of 0.0584. As for the last program offered at the bachelor level we have obtained a similarity of textual content between this program and the labor market requirements of 0.0519.

The same analysis has been done for the programs offered at the master's level, where even for these programs an almost textual similarity with the programs offered at the bachelor level has been acquired.

For the master's degree Computerized Automation and Robotics program, a textual similarity was obtained between this program and the textual content of the labor market requirements of 0.0513. Also for the Master's Computer Engineering program we have acquired a similarity between the content of this program and the labor market requirements of 0.0573. Whereas for the master's level Telecommunication program after analyzing through our model we have acquired a textural similarity between syllabus content and labor market requirements of 0.0511.

Since it is found on the website of the University of Pristina that PhD studies are offered, then such analysis has to be done for the studies offered at this level.

Compared to the studies offered at the two previous bachelor and master levels, PhD studies have a much smaller similarity. According to the analysis done by our automated model we have obtained a textural similarity between the syllabus content of this program and the textual content of the labor market requirements of 0.0120. As can be seen if compared to other levels we have a much smaller similarity. Also if compared to the similarity gained for the PhD program from South East European University we have a big difference. Of course, even for this part in the following chapters we will be able to provide the words that are missing in this textual content which results in us obtaining lower values of adaptation to the

demands of the labor market. In the following we will present the next analysis that has been done with the University of Tirana based on the programs offered by this University.

Also with this University, as with the two previous universities we are based on the information this university has published on its website. According to published data, this university has the smallest number of study programs in the field of technology at almost all levels.

4.8.3. University of Tirana versus market demands

In the following we will present the graph which contains the results of the comparison between the programs offered by this university and the labor market requirements.

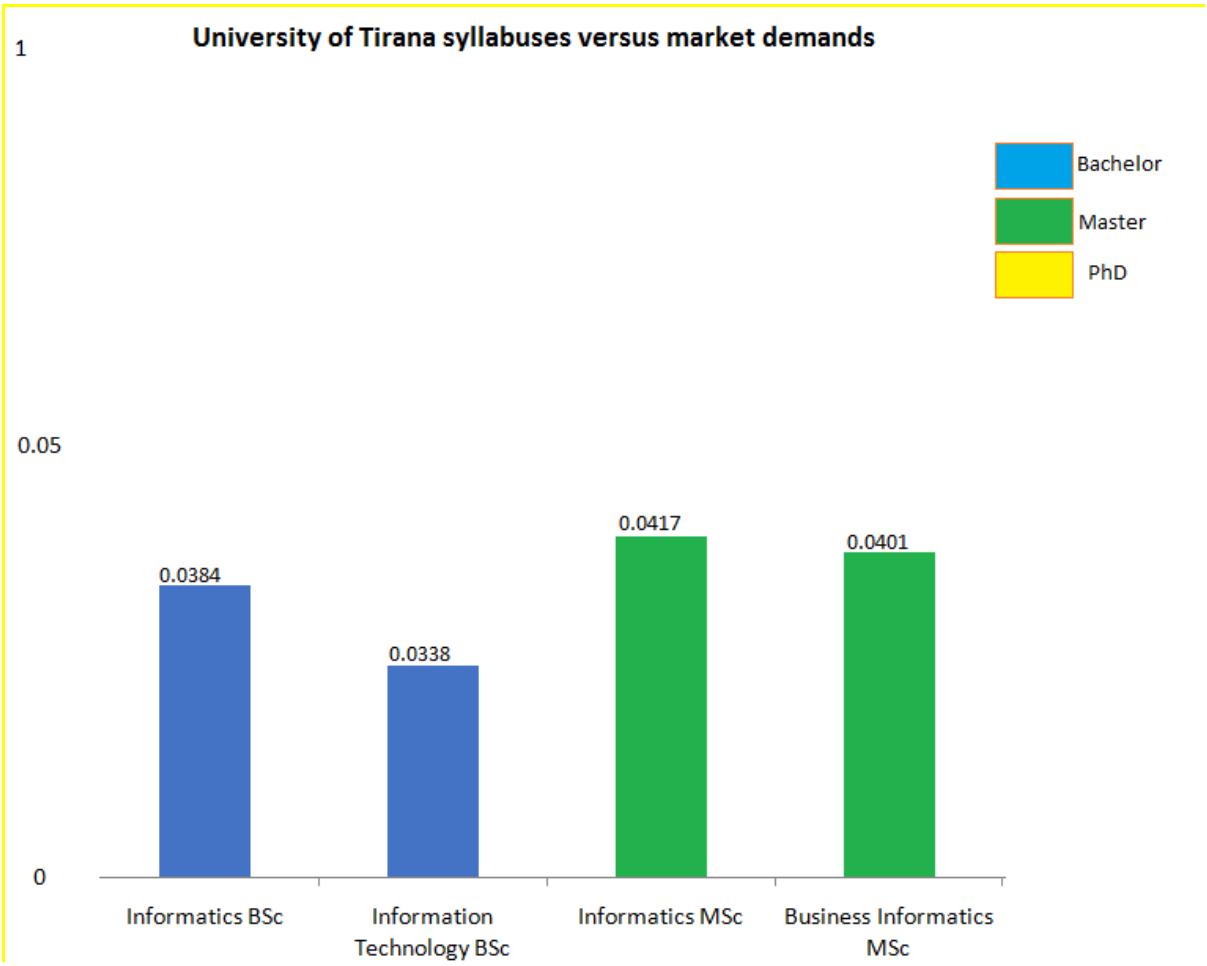


Figure 86. University of Tirana syllabuses versus market demands

As can be seen in figure 86, the number of programs published on the University of Tirana website is much smaller compared to South East European University and University of Pristina. However, we have made a comparison between the programs published by this university and the labor market requirements. The results obtained are much smaller

compared to the other two universities. As can be seen the analysis is done only for the two levels, master and bachelor since for the PhD level there is no official data on their website. The bachelor degree from this university offers the Informatics and Information Technology program. For the Informatics program after analyzing our model it turns out we have a similarity of textual content between the syllabus and the labor market requirements of 0.0384, which is a much smaller value than the other two universities in the region. Also for the Information Technology program we have a similarity of textual content between this program and the labor market requirements in the value of 0.0338.

The programs offered at the master's level stand out better than the programs offered at the bachelor's level, however they have a smaller fit than the two universities compared to the previous one.

As can be seen in figure 86, there are two programs offered at this level, Informatics and Business Informatics. Informatics software after analyzing our model turns out to have a textural similarity to content with labor market requirements of 0.0417. Whereas for the Business Informatics program we have a similarity of textual content between this program and the labor market requirements of 0.0401.

As we mentioned at the outset, if these programs are compared to the similarities we have acquired for South East European University and University of Pristina they result in much smaller similarities. If we make a more specific comparison between the Business Informatics programs, which are the same and offered by the two universities, there is a big difference. The following is a graphical comparison of the Business Informatics program.

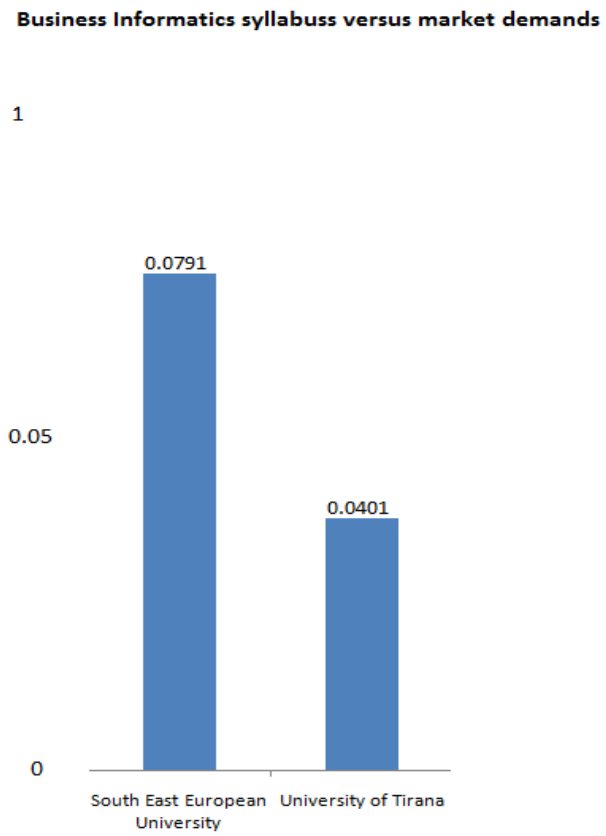


Figure 87. Business Informatics syllabus versus market demands

In figure 87 we present the comparison between South East European University and University of Tirana for the Business Informatics program because it is offered by both universities.

In figure 87 we can see that for Business Informatics program, South East European University has a similarity of 0.0791 and University of Tirana 0.0401. It turns out that there is a huge difference in the textual content of the program offered by the two universities. As with other analyzes, the Business Information Informatics program offered by the University of Tirana will include missing words in this corpus in order to provide recommendations on how to improve these syllabuses.

4.8.1. Universities versus European market demands

As we have a ready-to-read document containing the bulk of the bidding information in the European market, our model will be ready to analyze and compare it between the text. Below we will present the similar textual results between a study program and labor market offers.

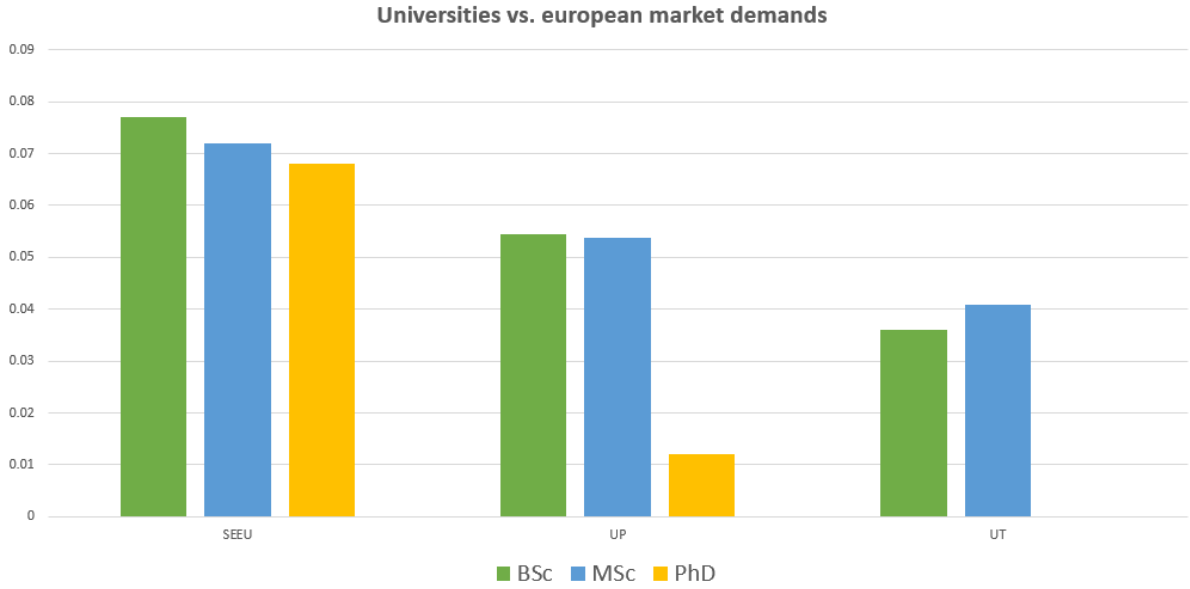


Figure 88. Universities vs European market demands

In Figure 88, the analysis of the study programs of the region's universities was presented in comparison to the bids of the field of technology. In the analysis, all the programs in the field of technology at all three levels of study were taken. In Figure 88 it can be seen that the similarity between texts between study programs and job offers in European countries is different. The South East European University has a major adaptation of the text of study and bidding programs compared to the University of Pristina and Tirana. The average Bachelor level for South East European University is approximately 0.08, for the master level of approximately 0.075 and for doctoral degrees of approximately 0.07

At the University of Pristina, the average of the textualization of study and bidding programs in European countries is about 0.055 for both Bachelor and Master levels and 0.015 for doctoral studies.

At the University of Tirana, the average for adapting the text of study and bidding programs to European countries is approximately 0.035 for the Bachelor level, and over 0.04 for the Master's level, while no PhD level has been achieved since there was a lack of syllabus the university website.

Therefore, based on the analysis applied by the automated model, the textualization of study programs and job offers for the European market is greater for South East European University compared to the other two Universities of the region.

4.8.2. Universities versus specific position vacancy

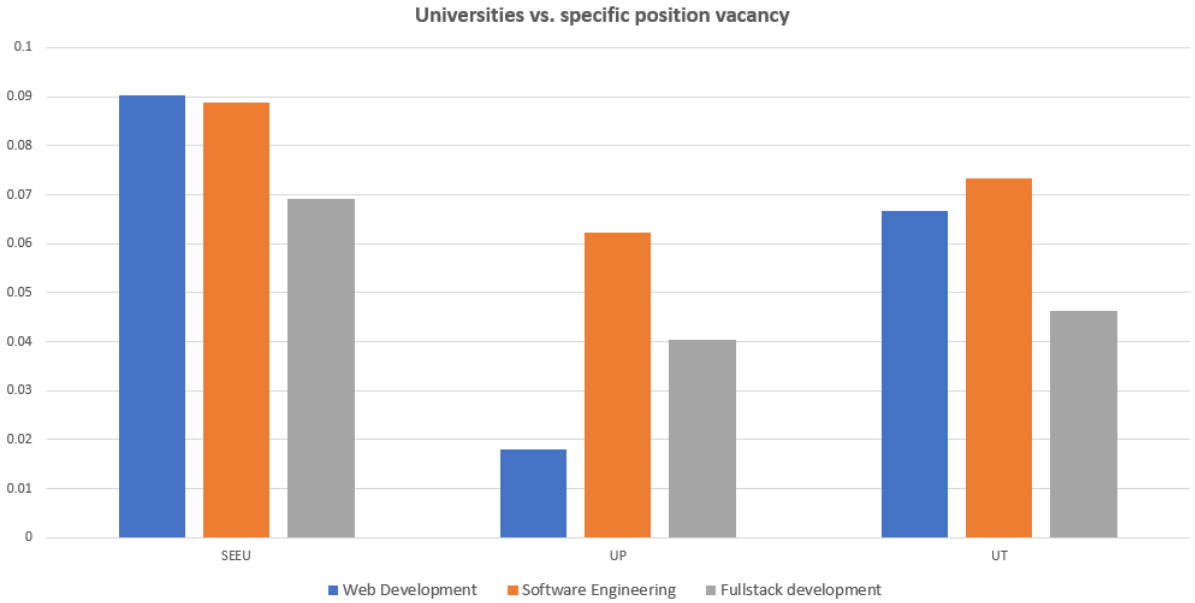


Figure 89. Universities vs specific market demand positions

Another analysis that is from our model is the analysis of study programs in the field of technology with specific positions of competitions in European countries. As can be, the analysis is divided into three competitions: Web development, Engineering software and Fullstack development. Three positions have been compared to curricula of universities in the fields of technology.

Based on Figure 89, it can be said that the South East European University has a twist of the text of study programs with three job offers at a level of 0.09 for Web development and engineering software, while for Fullstack development at a level of 0.07.

The University of Pristina has a smaller resemblance to the text of study and bidding programs. For the Web development postage, there is an adjustment of 0.02, for the engineering software there is an adjustment of 0.06, while for Fullstack development 0.04.

While the University of Tirana's web development site has a text alignment of 0.065, for engineering software 0.075, and for Fullstack Development there is an adaptation of 0.045.

Even in this analysis, we can point out that the biggest adaptation of the university curriculum texts to the bids of the field of technology is greatest for Southeast European University in comparison to the University of Pristina and Tirana.

4.8.3. Regional universities vs other universities

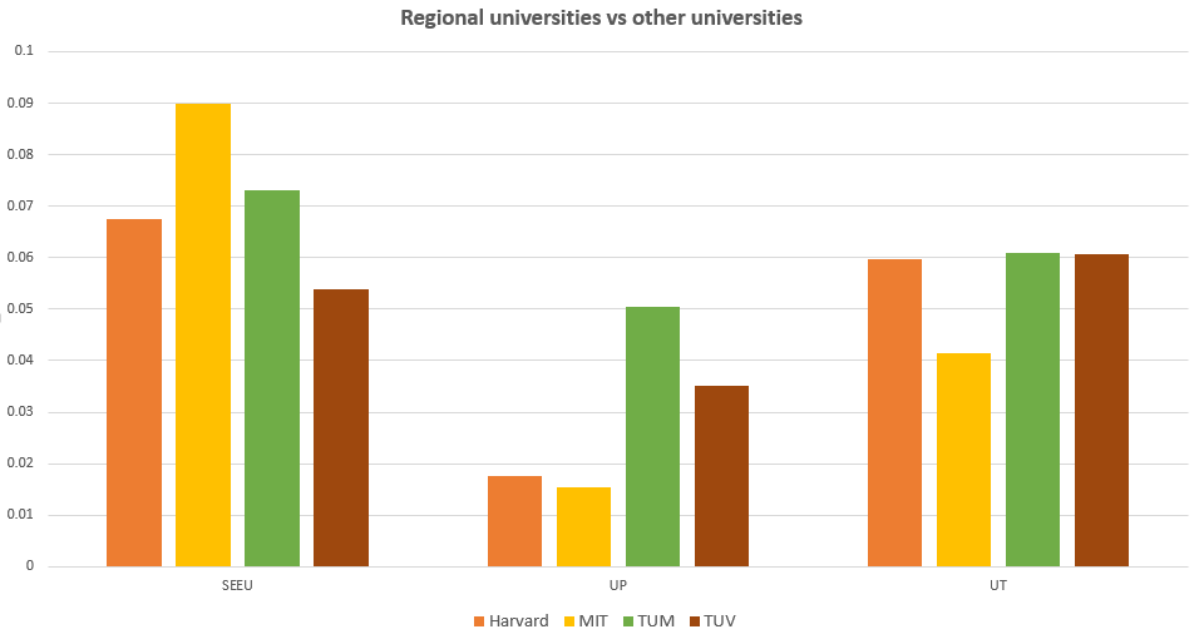


Figure 90. Regional universities programs vs top universities programs

The ranks analyzed which are applied are analysis of the curricula of the regional universities with the curricula of some top universities in different countries of the world. Part of the analysis are curricula in computer science at Harvard University and at MIT. While at the University of Munich, the curriculum of the Information Systems program is taken, and the Business Informatics curriculum for the University of Vienna is taken. All of these programs are compared with the study programs offered by the region's universities that have been part of your analysis.

In Figure 90, we can point out that the greatest extent of the Southeast European University curriculum is with MIT, then with München, continuing with Harvard, and the smallest of the adaptations of the Viennese at a level of 0.05. Compared to the South East European University, the University of Pristina has a much smaller version of the text of the study programs with the text of study programs of world universities. The minimum value for text adjustment is 0.01, while the maximum value is 0.05. While for Tirana University, the minimum value is 0.04, while the maximum value of the textualization of the study programs and texts of the study programs of the bows universities is 0.06.

Also based on the automated tailor-made analysis, the Southeast European University has a major adaptation of the textbook study texts of the study programs of European University programs.

4.8.4. Universities general comparison

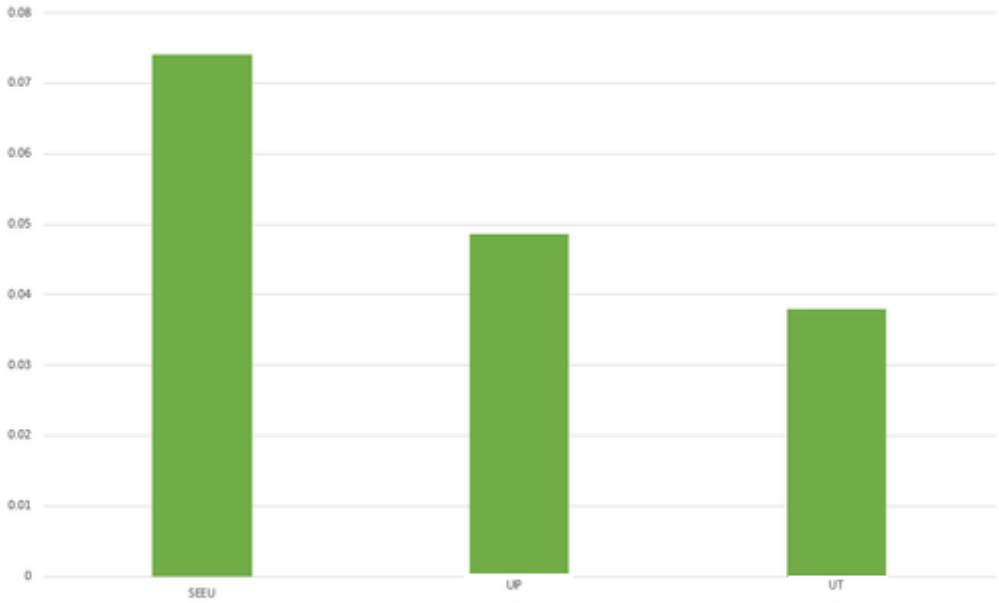


Figure 91. General comparisons of universities

Figure 91 shows the total value of all of the aforementioned comparisons. We have here comparisons with the European market, comparisons with specific positions, and comparisons with study programs in some universities in the world. As it may be, Southeast European University has a text alignment for all comparisons of 0.075. The University of Pristina has a text alignment for all comparisons of 0.04, and Tirana has a text alignment for all of the above mentioned comparisons of 0.05.

Therefore, as a conclusion, we can say that the South East European University has a much greater adaptation of the text in comparison with other universities in the region.

4.9. Conclusion

During the fourth chapter we have presented the form of the automated model algorithm that compares the university curriculum and the requirements of the labor market. We have initially illustrated the website identification form that publishes information on job offers in the field of technology. After identifying the bidding websites, it is the identification of the documents that hold the curricula for the universities. The explanation of the mathematical calculations applied in our model is their illustrations as mathematical calculations for later on with their application in python languages.

Since the implementation of all the algorithm components that enable the automated modeling work, we have been analyzing the study programs of the most recent bids in the field of technology. The universities we have taken as a case study are the University of: Southeast Europe, Pristina and Tirana. For all of these universities, the text of the study programs in the field of technology with the text of the bidding texts for European countries is compared. During our analysis we compare the text of study programs with job offers, specific offers and programs of some of the world's universities. After the analysis, we noticed that the University of Southeast Europe has a major adaptation of the text of the study programs for the supply of jobs to European countries. Apart analyzes with job offerings for European countries, South East European University has a great deal of textual and textual alignment with specific positions in the field of technology, and study programs of some of the world's universities in the field of technology.

PART 5

5. Evaluation of the model

During the fifth chapter we will talk about evaluation of the model where as part of it will be the stability of the model, where we will test how stable our model is. The second part will be clustering vacancy corpus. We will split the model into clusters and make comparisons between syllabuses and labor market requirements with each cluster. In the third part we will do silhouette analysis where we will find the appropriate number of clusters of our system. Finally, the fourth section will include the vacancy corpus analysis where the frequency of the most frequently mentioned words and the least frequently mentioned words will be presented. Also in this section will be in-depth analysis of the words that are mentioned in job offers.

5.1. Stability of the model

In order to test the stability of the model we will do some testing to arrive at definitive conclusions about whether our model is stable or not. The tests that will be done on our system are:

- ***Removing stop words and comparing with previous results.***
- ***Removing some job offers.***
- ***Increasing the volume of corpus by adding new job offer.***
- ***Cross-validation.***
- ***Comparison with other models.***

We will first start by removing the stop words from our system and compare with the previous results and present the results that are achieved in order to see if there is a big difference and what is the importance of stop words.

The next step is to remove some of the contests from our corpus and compare it with the previous results to confirm the stability of our model. According to our expectations, even

after removing or adding competitions to our database, the similarity results should not change much.

The next analysis to test the robustness of the system is the addition of some competitions to our database, where we also expect that the results will change most as new competitions are expected to be added in the future that will increase the volume of our corpus. .

And the last step to test the stability of our model is cross-valiation with the data in the competition corpus. The entire competition corpus will be divided into different sections then comparisons will be made by removing each section from the corpus. In the following we will present the results for each of the above analysis.

5.1.1. Removing stop words

As mentioned above, in order to test the stability of our model one of the steps is to remove stop words from the body of the competition data. Of course, these words should not be overweight, and based on the experiments we mentioned in chapter four, these words will even lower their weight if they are in the corps.

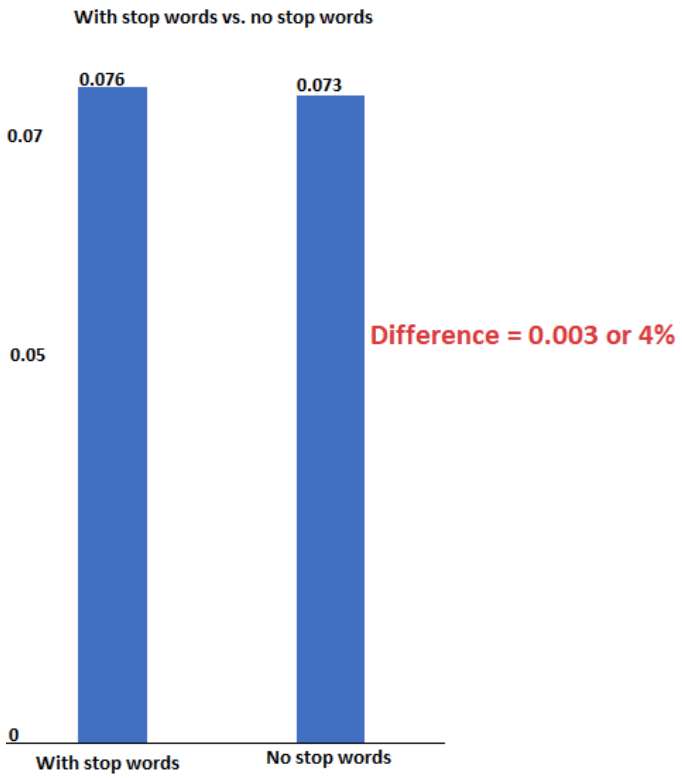


Figure 92. Difference of corpus with stop words and with no stop words

Figure 92 shows the similarity between labor market requirements and university curricula with stop words and with no wop words. As can be observed when a random comparison is made between labor market requirements and a randomly selected syllabus, the similarity of the textual content acquired is 0.076. By means of an algorithm we removed the stop words from the corpus containing the data on labor market requirements, and again calculated the similarity between the textual content. After removing the stop words from the labor market requirements corpus we obtained a textural similarity between the labor market requirements and the university curriculum of 0.073. As can be seen in figure 92, the difference between stop words and no stop words is 0.003 or 4%. From this it can be concluded that the importance of stop words is not too great as normalization techniques have been applied in our model which reduce the weight of the most frequently used words. In the following we will present the results as we remove from our data a portion of the contests.

5.1.2. Removing job vacancies

The content of the corpus with the demands of the labor market is of great importance, but more important is the processing of its content. After processing the labor market requirements corpus, we have compared the complete content and part of the contests.
Figure 93 shows the difference of textual similarity between labor market requirements and a randomly selected syllabus. The difference is between the complete content of the corpus with labor market requirements data, and the removal of a significant portion of the content contests. As can be seen the result of the textual similarity between the full content of the labor market requirements and the syllabus is 0.076. After removing a significant portion of the textual content of the labor market corpus, the technique of comparing textual similarity was again applied, and the result obtained was 0.072. As can be seen in figure 93 the difference is very small since we have a difference of 0.004 or if we calculate in percentage we have a difference of 5%.

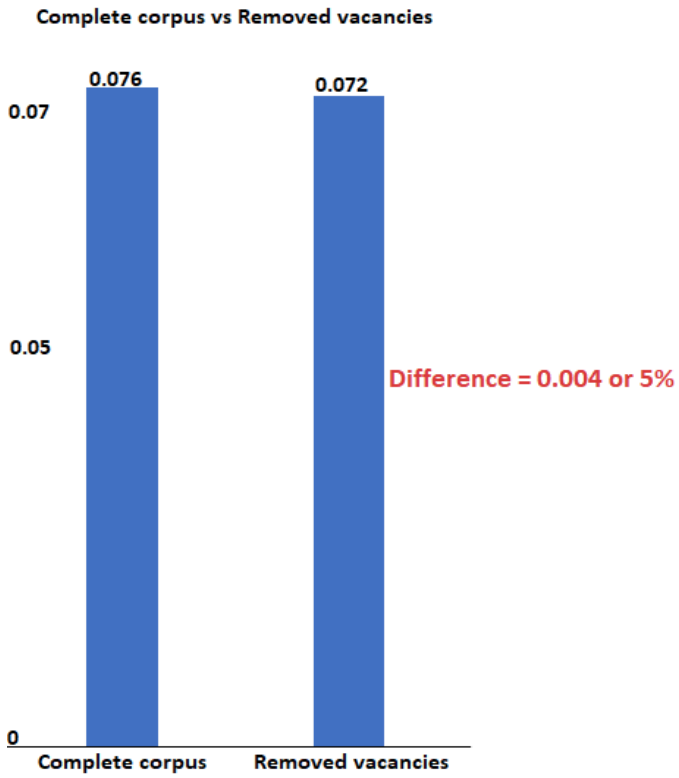


Figure 93. Difference of complete corpus and removed some vacancies corpus

Even in this case the difference is very small and this gives great importance to our model since despite the changes that the labor market may have for a short time, the results will not change much, and of course this makes the model very stable. In the following we will present the case when we add some new competitions to the labor market demand corps, which further supports the sustainability of our model.

5.1.3. Increasing the volume of corpus by adding new job offer.

As mentioned in the chapters above, in the coming years, there is expected to be a large increase in labor market demands in the field of technology. What is important about our model is that it is applicable even after a time when there are new competitions published in the field of technology. The next step that will test the viability of our model is to increase the volume of the corpus with data on technology contests.

The next analysis is done again comparing the complete content of the corpus with the labor market requirements with a randomly selected syllabus. After this analysis, the volume has

been increased with new contests in our corps, and again the comparison of textual similarity has been made.

New competitions have also been taken from the website which was initially used to launch the published competitions, but also new competitions from other websites which are published in the field of technology.

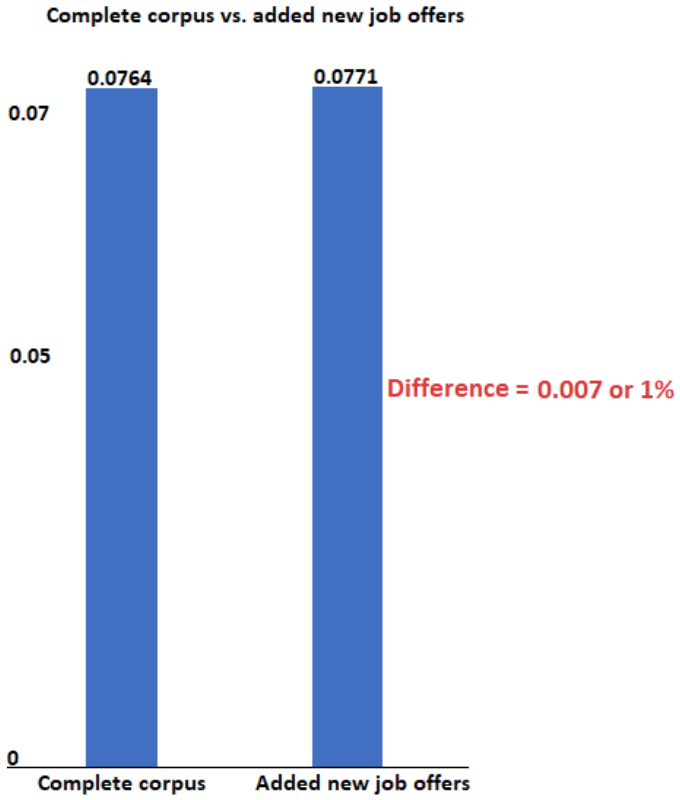


Figure 94. Difference between complete corpus and added new job offers corpus

Figure 94 shows the comparison of the similarity of textual content between the corpus created by us, and the corpus after we added some new contests. According to previous analyzes and the results presented in figure 94, the similarity of textual content between labor market requirements and a randomly selected syllabus is 0.0764. Whereas, after adding some new job offers to our corpus this result has changed to 0.007, and we have obtained a textural similarity score between the labor market requirements and the university curriculum of 0.0771. This difference is very small as it is only 1%, and by calculating this small difference we can conclude that our system is very stable in terms of comparing labor market requirements and curricula offered by the university. Also this change is of great importance for the fact that almost every day new competitions are published on the website, and those new competitions will not affect us to get results different from those we have concluded in

our model. In the following we will present the case of cross validation, which is the fifth step of testing our system stability.

5.1.4. Cross validation

The next step that confirms the stability of our system is cross validation with the data we have in our body. The way we are going to do this analysis is by dividing our competition corpus into ten different pieces, then we will make a text mix and remove 10% of each piece from the text. First remove the first 10%, compare, then place the first 10% and remove the second 10% and so on.

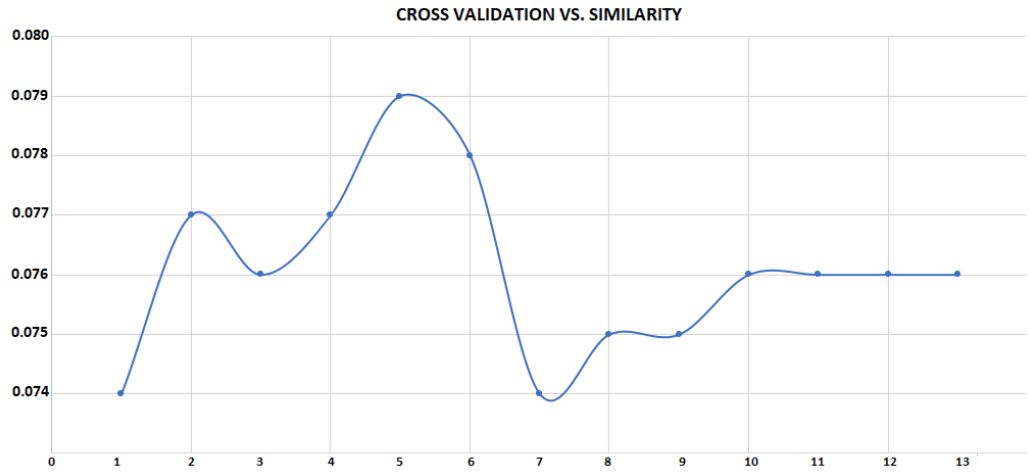


Figure 95. Cross validation of vacancy corpus

Figure 95 shows the cross validation analysis of the job vacancies corps and the placement of new vacancies in our corps. As mentioned above, and as it is known that cross validation analysis works, our body is divided into ten parts. After dividing the competition corpus into ten parts, the text was mixed, and in this way we removed 10% of the corpus for each part.
As can be seen in figure 95, when we removed the first 10% we obtained a result of textual similarity between the labor market requirements and university curricula of 0.074. After this step we returned the first 10% to the corpus again, and removed the second 10% from the corpus, and after the comparison we obtained a similarity score of 0.077. The third part has a score of 0.076, with the fourth part we have 0.077. We have seen a sharp increase in resemblance since the fifth part of our corps was removed, where we gained a similarity of 0.079, and with the removal of the sixth part we gained a similarity of 0.078. After the seventh part is removed, we have a sharp decline as the likelihood falls back to 0.074, and with the eighth and ninth part we have a stabilization of results as the level of textual similarity

increases to 0.075 for the eighth part, and 0.076 for the second part. nine. As can be seen with the completed corps we have a similarity level of 0.076, and this result begins to stabilize as we begin to place new contests in our corps. As can be seen in Figure 95, with new competition placements, where the graphs show as the eleventh, twelfth, and thirteenth we have almost the same results as the completed corps with a small margin of 1%. So such an analysis supported all the preliminary analysis which was done in order to verify the stability of our system. The next step that underpins the consistency of our model is to compare our model with another model that makes textual content comparisons.

5.1.5. Comparison with other models.

Comparison with an existing model which compares textual similarity between different documents is of great importance in the sustainability of our model. In the following we will present the differences in results from our model and from another model available online, as well as cross validation analysis of the other model.

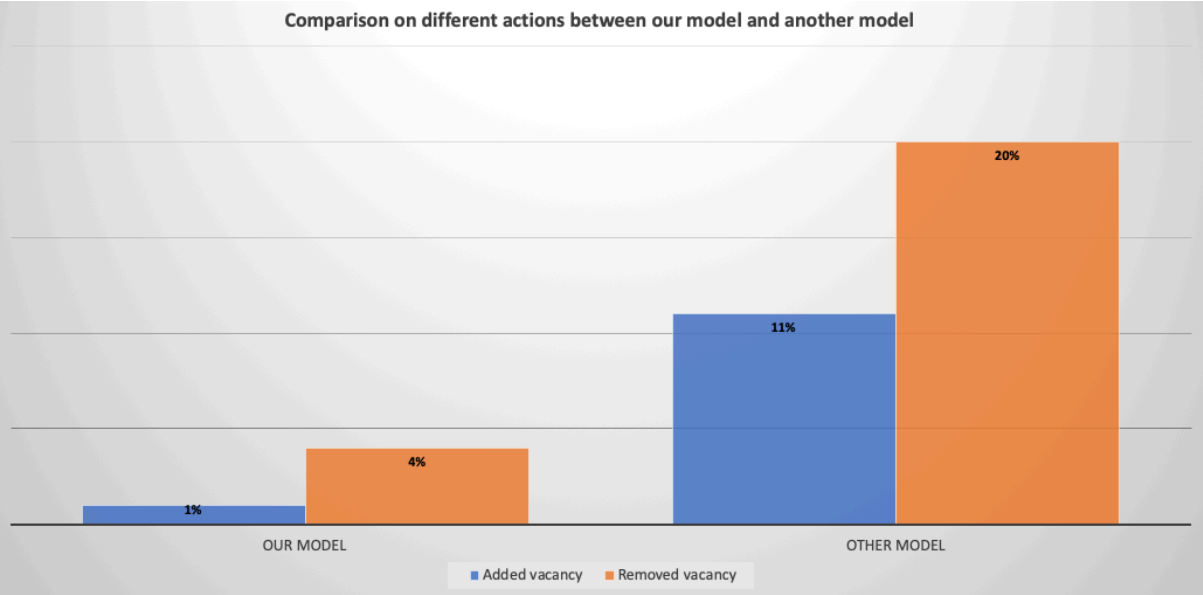


Figure 96. Comparison on different actions between our model and another model

Figure 96 shows a comparison of our model with another model which also compares the textual content between different documents. Unlike our model, the other model does not use data normalization, which negatively affects the outcome that the model ultimately achieves.

According to the analysis, our model provides accurate and small changes as shown in the graphs above. As can be seen in figure 96, our model after the addition of new contests in the competition corps gives us a score of only 1% difference compared to the corpus we own. Whereas after leaving the contests from our corps, the model gives us a result which is only 4% difference with the result that the model gives with the original corps.

While the other model that we have compared to ours, at this point it yields results that have a big difference with the first results after comparing textual content between labor market requirements and university curricula.

As can be seen in figure 96, after adding new contests to our corps, the other model yields a score that is 11% higher than the primary scores the model offers with the original corps. Whereas after removing some contests from our corps, the other model yields a score that is 20% different from the primary scores the model offers with the original corps. Of course, even at this point our model has an advantage in terms of consistency and accuracy, since text normalization methods have been applied in our model which influence the model to provide more accurate results. In the following we will present the cross validation analysis of the other model, to see how much the model has stability compared to our model.

5.1.6. Cross validation of other model

Another analysis that compares our model with another model that compares textual content is cross validation analysis as done in our model. Surely such an analysis will greatly help us to know how consistent and accurate our model is. The analysis will be the same as in our model, where the corpus will be divided into 10 parts, the textual mix will be removed, and 10% of the corpus will be removed until complete corpus content is reached. Once we get to the full content of the corps we will add new contests to see if the system succeeds in stabilizing the results the same as our model.

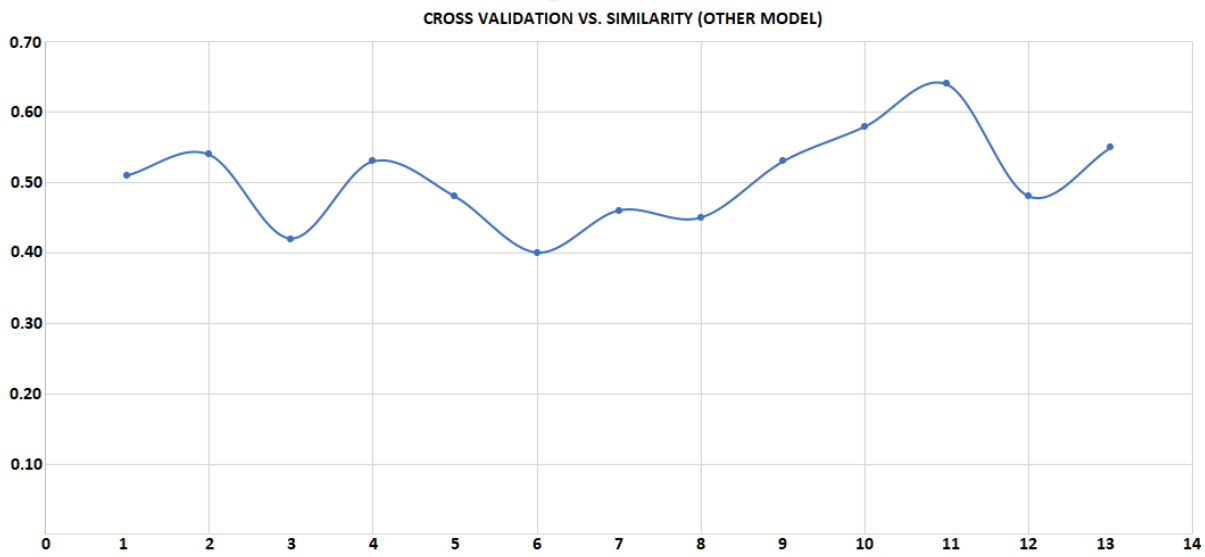


Figure 97. Cross validation of vacancy corpus (other model)

Figure 97 presents the cross validation analysis of the vacancy corpus of the other model. The same analysis was applied as with our model. We initially split the corpus into 10 parts, and we did a text mix to continue with each section later. As can be seen in Figure 97, the other model has a different level of measurement of textual similarity between different documents. After we removed the first part the similarity level of the other model is 0.50. We reset the first part, and removed the second part and the similarity level reached 0.55. Again we continued with all parts until we reached the bottom of the tenth, where all our corpus is with job vacancies, and the textual similarity between the labor market requirements and the syllabus at this point was reached at 0.58. As can be seen when the placement of new competitions is made the model has not stabilized, but again we have drastic differences in textual similarity between the two documents.

Certainly comparing our model with such an existing model makes our model much more consistent and accurate, since the existing model when we add new competitions that are published, then the results vary by 10-20%. as shown in the earlier chapters. Therefore we can conclude that our model provides accurate and consistent data compared to other models that make comparisons of textual content between different documents.

5.2. Clustering evaluation

Before moving on to clustering our corps, we must first analyze our corps in order to determine the optimal or most appropriate number of clusters that our corps will contain.

Since in our model we apply the k-means clustering method, where we are required to determine the number of clusters to be created, then we still need to perform an analysis in order to determine the optimal number of clusters to be created. created.

There are a number of methods used to analyze clusters and to determine the optimal number of clusters, and of course none yields the same results as it depends on the method used. The methods used to analyze the optimal number of clusters are:

- ***Silhouette method.***
- ***Elbow method.***
- ***R analysis.***
- ***Gap statistic method.***

All of the methods mentioned above are methods used to analyze the optimal number of clusters. Of course there is no single method used to analyze the cluster, but it depends on the form of the data and the method where to apply which one is more accurate and which is more appropriate.

Importantly, all of the above mentioned methods are used to determine the optimal number of clusters and some of them are also used for statistical research.

In our case we will use the silhouette method as the most widespread method in order to determine the optimal number of clusters we will divide into our corpus. Later these clusters will be compared to each other, but will also be compared to the region's university syllabuses in order to determine which cluster group is more closely related to the textual content of the university syllabus.

Once the syllabus which has greater and lesser textual similarity between university curricula and labor market requirements has been identified, the analysis of the words contained in that cluster will be done. Based on these words that we identify to be part of the cluster, we will be able to make recommendations on which words should be included in the university syllabus in order for the syllabus to be more in line with labor market requirements. .

Of course, such an analysis will be of great help and support to us, as it will validate the analyzes we have previously made between the region's university syllabuses and labor market requirements.

5.2.1. Silhouette analysis

The method which shows how close an object is to its cluster compared to the other cluster is known as Silhouette analysis. According to Kaufman and Rousseeuw 1990, the average obtained by silhouette analysis shows exactly how optimal the number of clusters created is. A higher average indicates that the number k of the clusters is optimal and at such numbers it is preferable to divide the corpus with textual content.

The values that can be obtained after applying Silhouette analysis are from -1 to +1. The higher the value, the closer the object is to its cluster, and vice versa, the smaller the value that is acquired, the farther away is the object with its cluster. After calculating these values, an average is obtained which shows the optimal number of clusters. This number is very easy to assign, since the number where we get the highest average is the optimal number that our cluster will contain.

Whether or not an object is aligned with its cluster can be measured in several forms, but in the case of silhouette analysis this is done using the Euclidean distance method. As we mentioned in the previous chapters, Euclidean distance works by calculating the two points that are placed in Euclidean space.

Since in our case we are dealing with textual data, as previously mentioned as a cluster method we will use the k-means method. In the following we will present the mathematical calculations that are used to perform the Silhouette analysis to proceed later with its application through algorithms and the Python language in order to perform the optimal number calculations of our corpus.

According to (Rodriguez, 2019), we present a case where we have created some clusters which we will present with CM and CN, and compare the distance a between Oi and other objects in CM, and the distance b to between Oi and other objects in CN.

$$a(O_i) = \frac{1}{|C_M| - 1} \sum_{O_j \in C_M, O_j \neq O_i} d(O_i, O_j)$$

$$b(O_i) = minC_N \neq C_M \frac{1}{|C_B|} \sum_{O_j \in C_N} d(O_i, O_j)$$

$$silhouette(O_i) = \frac{b(O_i) - a(O_i)}{\max\{a(O_i), b(O_i)\}}$$

In the above equations, the mathematical equations which calculate the average silhouette are presented. As we can see, three equations are presented which contain the calculation steps, starting from the first step presented in the first equation. According to this equation, a (Oi) is equal to the division between 1 and the absolute value of the cluster CM minus 1 and the sum of the distance between Oi and Oj where both are objects of a cluster, but must not be equal to each other.

Once the first cluster is computed, we must define the second cluster that in our case we define with b(Oi). This cluster is the minimum distance of one of the objects of the first cluster, but it must never be part of the first cluster. As we can see in the second equation, it is equal to the minimum of the first cluster that is different from the second cluster. Then this minimum value is multiplied by 1 partition for the absolute value of CN which in this case is the second cluster. And this value is also reduced by the sum of the distance between the objects Oi and Oj, where Oj is an element of the second CN cluster.

Once the second cluster is defined, then in the third equation we do the silhouette calculation.

According to the third equation, the silhouette equals the division between the subtraction of b(Oi) and a(Oi) and the maximum value between the first cluster a (Oi) and the second cluster b(Oi).

If we want to calculate the classification quality of a single object, then we can extend the last silhouette equation. Below we will present the case of calculating the quality of all objects that are part of a cluster, as well as the case of calculating the quality of clusters one by one.

$$Silhouette(C_i) = \frac{1}{|C_i|} \sum_{O_j \in C_i} silhouette(O_j)$$

This equation presents the case of calculating the quality of all objects that are part of a cluster. As can be seen in the equation, the silhouette equals the division of the value 1 and Ci, and the multiplication of this value by the sum of the silhouette objects (Oj), where the object (Oj) must be an element of the first cluster. While calculating the quality value for each cluster according to the equation below.

$$Silhouette(C) = \overline{silhouette(m)} = \frac{1}{m} \sum_{i=1}^{m} silhouette\ (C_i)$$

In the above equation an equation is presented which calculates the quality value for each cluster individually. As we can see, silhouette (C) is equal to silhouette (m) which is a vinculum, that is, the set of all cluster values. And this value is equal to the division between 1 and m cluster, and the output of this value is the sum of the silhouette (Ci), where i starts from 1 to m which is the number of clusters that are defined in the system .

So, as can be seen, the calculation of the quality of the cluster and the objects that are part of the cluster, through silhouette analysis, can be done in a very precise way through the mathematical calculations as mentioned earlier presented by Kaufman and Rousseeuw. After the mathematical equations that make up the silhouette are presented, we will now present this analysis through the algorithm which will be constructed in the Python language. Once we have created the algorithm, we will apply it to our corpus, to see what is the optimal number of clusters we need to create in our system, to proceed later with comparing syllabuses with each cluster, and comparing all clusters with each other. The comparison of these clusters will be made using our model in order to determine the similarity of textual content between these clusters.

```
import numpy as np
import matplotlib.pyplot as plt
import python_utilities
from sklearn import metrics
from sklearn.cluster import KMeans

import utilities

```

Figure 98. Importation of libraries for silhouette analysis

Figure 98 shows the part of the bookstore import we need in order to implement silhouette analysis. As can be seen, there are all the libraries that are needed starting from the mathematical calculations that the system will do to those needed to create the figures with the results obtained.

```
#Load data
data = open('/Users/Ylber/Desktop/cluster 0.txt', 'r').readlines()

scores = []
range_values = np.arange(2, 10)
```

Figure 99. Load data of vacancy corpus

The second part is the load data of vacancy corpus, where in our case it is the data that has been processed and prepared for this part. The data to be imported is the data that was originally converted to vector values. After this part, it will be the model training part.

```
for i in range_values:
    #Train the model
    kmeans = KMeans(init='k-means++', n_clusters=i, n_init=10)
    kmeans.fit(data)
    score = metrics.silhouette_score(data, kmeanslabels_,
                metric='euclidean', sample_size=len(data))

    print('\nNumber of clusters =', i)
    print('Silhouette score =', score)

    scores.append(score)
```

Figure 100. Train model of silhouette analysis

Figure 100 shows the training part of our model which will do the silhouette analysis. As can be seen in figure 100, this section defines the type of clusters that in this case is k-means, as well as the definition of the score in our case is silhouette, and the measurement of the distance between objects that are within clusters will be done with Euclidean distance.

After defining the clusters and the score, the results are finally printed in textual form, such as the number of clusters, as well as the silhouette score which are defined above.

```
    #Plot scores
plt.figure()
plt.bar(range_values, scores, width=0.6, color='k', align='center')
plt.title('Silhouette score vs number of clusters')
plt.show()

    #Plot data
plt.figure()
plt.scatter(data[:,0], data[:,1], color='k', s=30, marker='o', facecolors='none')
x_min, x_max = min(data[:, 0]) - 1, max(data[:, 0]) + 1
y_min, y_max = min(data[:, 0]) - 1, max(data[:, 0]) + 1
plt.title('Input data')
plt.xlim(x_min, x_max)
plt.ylim(y_min, y_max)
plt.xticks(())
plt.yticks(())

plt.show()
```

Figure 101. Plot data and plot scores definition

The last part, which is part of the silhouette score calculation algorithm, is the creation of figures that make the visual representation of the silhouette score we obtain. These figures are created through the libraries we created at the beginning of our algorithm, and as can be seen in the first section, bar charts are created that represent the relationship between the silhouette score and the number of clusters. This section also shows the optimal number of clusters in graphical form, where we can see what is the result of all the clusters created by our system. In our case, this model will be able to graphically represent the optimal number of clusters created by the system, to continue with further analysis with these clusters.

In addition to showing the bar charts between the silhouette score and the number of clusters, this algorithm will also display the objects of the clusters that are part of each cluster. Here one can see which cluster is closest to the other, as well as the distance of each object which in this case is calculated by Euclidean distance. After defining the algorithm, and preparing the data, then we are ready to execute the algorithm which we will present using the figures below.

Figure 102. Silhouette score versus number of clusters

Figure 102 shows the number of clusters relative to the silhouette score that each cluster has. As we can see in this graph, the number of clusters is from 2 to 10. The smallest silhouette score reached by the cluster set is 0.27 where we have two clusters. Then the value of silhouette score for 3 clusters is 0.36, for 4 clusters we have the value of 0.41. The highest value is with 5 clusters, where the silhouette score reaches 0.78, which represents the optimal number of clusters that our research should contain. For 6 clusters we have a decrease in silhouette score, where its value reaches 0.48, and for 7 clusters we have 0.49. As for the last two groups with 9 clusters and 10 clusters we have values of 0.44 and 0.34.

Such an analysis helps us very much to determine the number of clusters that our research will contain. According to this analysis, our labor market demand corpus will contain 5 clusters, as it is the highest value of the silhouette score which is achieved by our model.

Since we have the optimal number of clusters, we will then use these 5 clusters to compare them with the university syllabuses that are part of the analysis, but also to compare the textual similarity between all the clusters created. Below we will present the graph of the full score silhouette, where based on it we will be able to see which cluster is closer to each other, and which cluster objects are more aligned with other cluster objects.

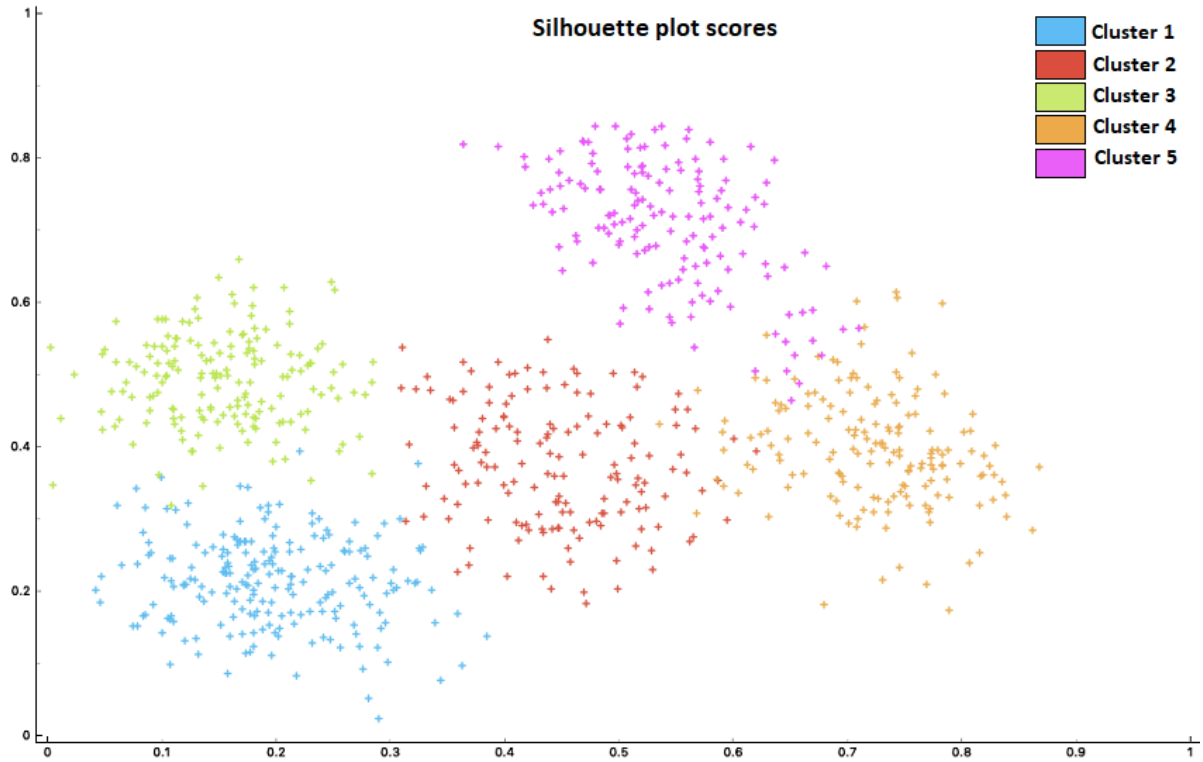


Figure 103. Silhouette plot score

Figure 103 shows the full score silhouette after running the algorithm we built for such an analysis. As can be seen in Figure 103, only the optimal number of clusters defined by the system is shown here. So we have 5 clusters that are presented in the x and y dimensions, where we can even guess which one is closest to the other. According to the graph we have the color split of all clusters, from 1 to 5, and we can also see which cluster is closest to the other. According to the graph we have a proximity between the objects of cluster 1 and cluster 2, since the distance between them is very small, but they are not equal to each other. There is also a small distance between cluster 1 and cluster 3, and between cluster 2 and cluster 3. We also have a small distance between cluster 2 and cluster 5, as well as cluster 3 and cluster 5, while there are large distances between cluster 5 and cluster 1 objects, as well as cluster 5 and cluster 2. According to the graph that made the representation of objects between clusters, there is no proximity between cluster 5 objects. and cluster 1, and cluster 5 and cluster 3. In the next chapter we will present the clusters created by the system to proceed with further analysis which validate our data accurately and efficiently.

5.3. Clustering vacancy corpus

The inclusion of clustering in our research is very necessary as it helps us to identify the cluster of similarity words found in the syllabuses offered by universities in the field of technology. Since in the previous chapter the analysis of the optimal number of clusters was done, then we have ready the number of clusters that will be used in our analysis. As shown in the graphs above, the number of clusters that will be part of our analysis will be 5, and each of them will be compared to the syllabus to find which words have the most similarity. , and what are the words that have less similarity or are less mentioned in the university syllabus. In the following we will present graphs of comparison of each cluster with university syllabuses.

5.3.1. South East European University versus clusters

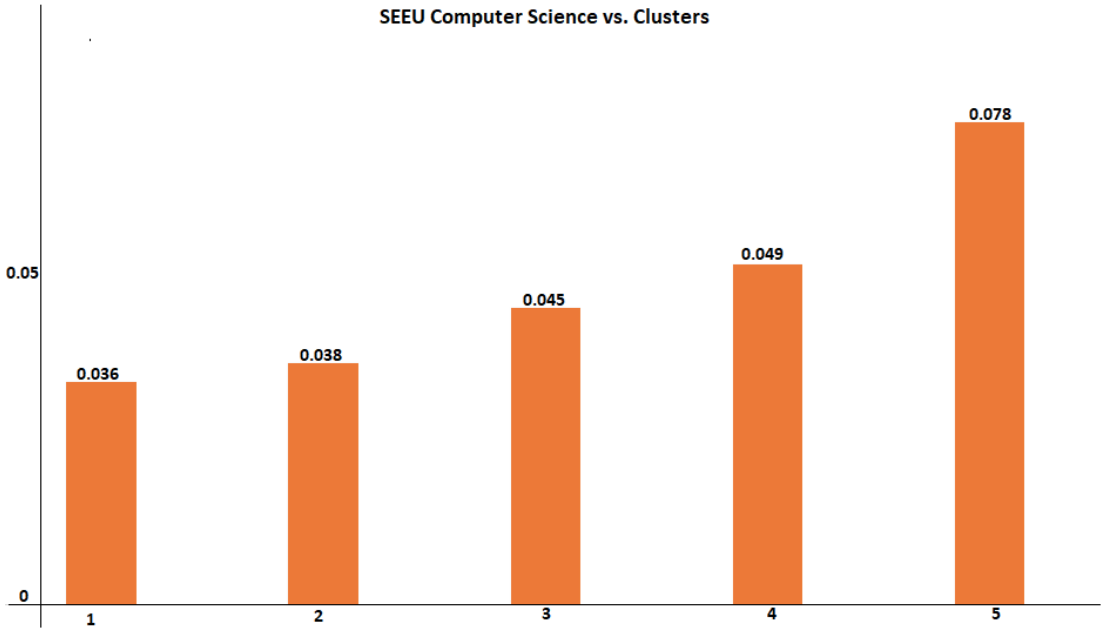


Figure 104. South East European University versus clusters

Figure 104 shows the relationship between clusters and similarity with the Computer Science syllabus of South East European University. As we can see, the textural similarity between the clusters and the syllabus ranges from 0.036 to 0.078. With the first cluster we have a similarity of 0.036, with the second cluster we have a similarity of 0.038. We have a greater similarity with the other two clusters since we have similarities of 0.045 and 0.049. And the biggest similarity is achieved with the fifth cluster where we have a textual content similarity of 0.078.

In the following we will present the analysis that has been made between the University of Pristina and the clusters created by the system.

5.3.2. University of Pristina versus clusters

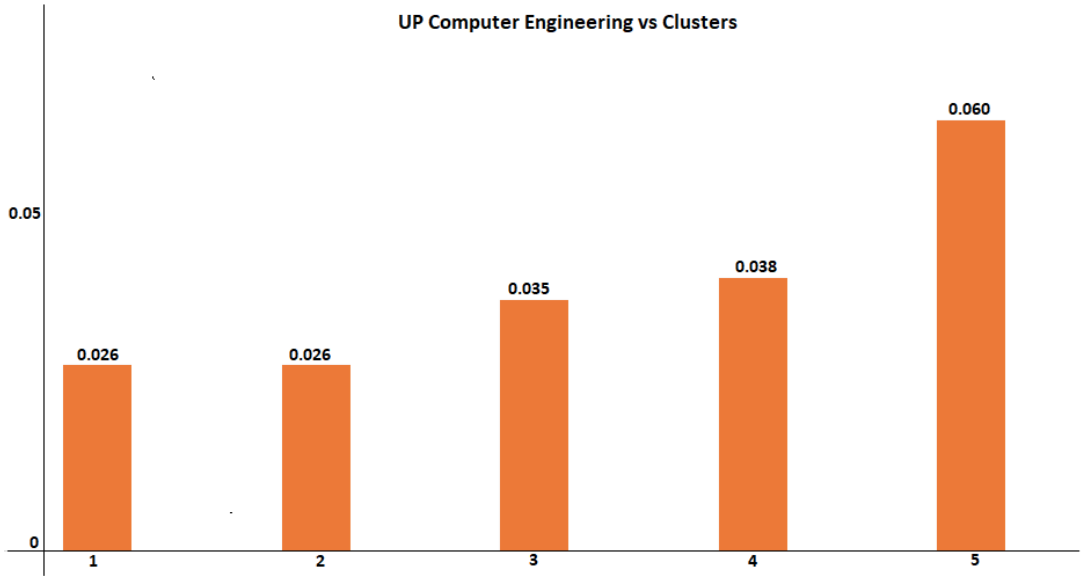


Figure 105. University of Pristina versus clusters

Figure 105 shows the analysis between the University of Pristina Computer Engineering syllabus and the clusters created by the system. In this analysis we have lower results than South East European University, since textural similarity ranges from 0.026 to 0.060 as the highest value. As we can see in figure 105, with the first cluster we have a textural similarity of 0.026, also with the second cluster we have a similarity of 0.026. A greater similarity is achieved with the third and fourth clusters, since we have a similarity of 0.035 and 0.038. And the maximum value reached between the University of Pristina syllabus and clusters is 0.060 with the fifth cluster created by our system.

As mentioned at the outset, at the University of Pristina we have a smaller textual similarity between the Computer Engineering syllabus and the clusters created by the system. Certainly such an analysis is supported by the results of the above analysis which also make the University of Pristina the second university that has similarity of textual content to the demands of the labor market. In the following we will present the analysis made between the University of Tirana and the clusters created by our system.

5.3.3. University of Tirana versus clusters

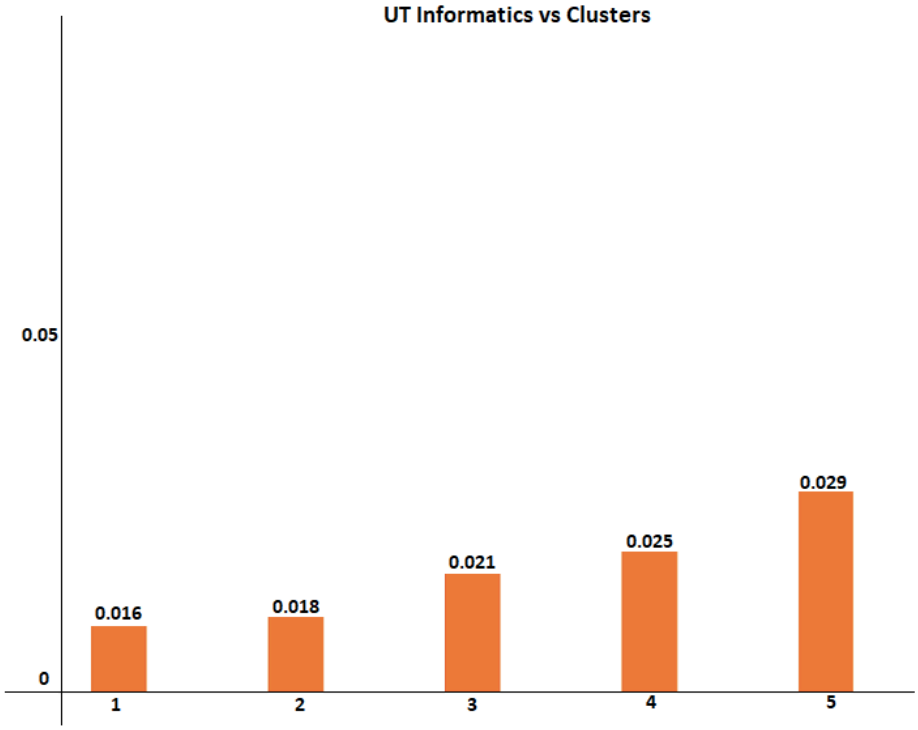


Figure 106. University of Tirana versus clusters

As for the two universities in the region, the same analysis was done for the University of Tirana, where the textual similarity between the Informatics syllabus and the clusters created by our system was compared. As can be seen in Figure 106, the similarity is very small compared to the two previous universities. The smallest value of textural similarity is 0.016 with the first cluster. We have a similarity of 0.018 with the second cluster, and we have a similarity of 0.025 with the third cluster. The maximum similarity is also with the fifth cluster, since the similarity value between the Informatics syllabus and the fifth cluster is 0.029.

Compared to the other two universities, the similarity of the University of Tirana is very small, and it ranks as the third university for the similarity of textual content it has with the labor market requirements. Of course, even to this analysis we have a great deal of support, as it supports all the analyzes that have been done before, and that all have put the University of Tirana as the third university. In the following we will present a graph for all three universities.

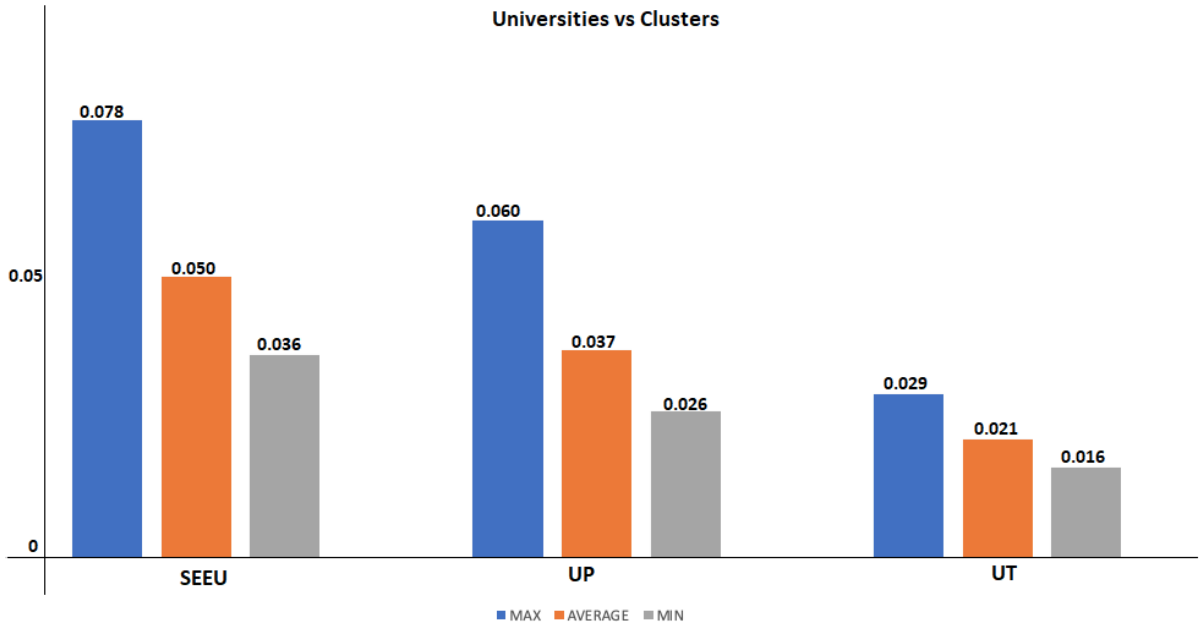


Figure 107. Universities versus clusters

Figure 107 shows a graph showing the maximum, average, and minimum values that universities have with the clusters created by the system.

Based on the graph, we can see that the maximum value for South East European University is 0.078, the average value is 0.050 and the minimum value for textural similarity is 0.036.

For the University of Pristina, the maximum value of textual similarity is 0.060, the average value is 0.037 and the minimum value of textual similarity is 0.026.

For the University of Tirana, the maximum value is 0.029, the average value is 0.021, and the minimum value for textual similarity is 0.016.

At all three universities, maximum similarity was achieved with cluster five, and minimum similarity was achieved with cluster one. Based on these values, we will make an analysis of these clusters about their textual content to analyze which words are the most repeated, and which are the least repeated. Certainly it is the number of syllabuses that are repeated in syllabuses that affect the small similarity or high similarity of textual content to the demands of the labor market. In the following we will present the contents of cluster 1 and cluster 5 words to see which syllabus words are similar and which words are not.

5.3.4. Most similar cluster with syllabuses

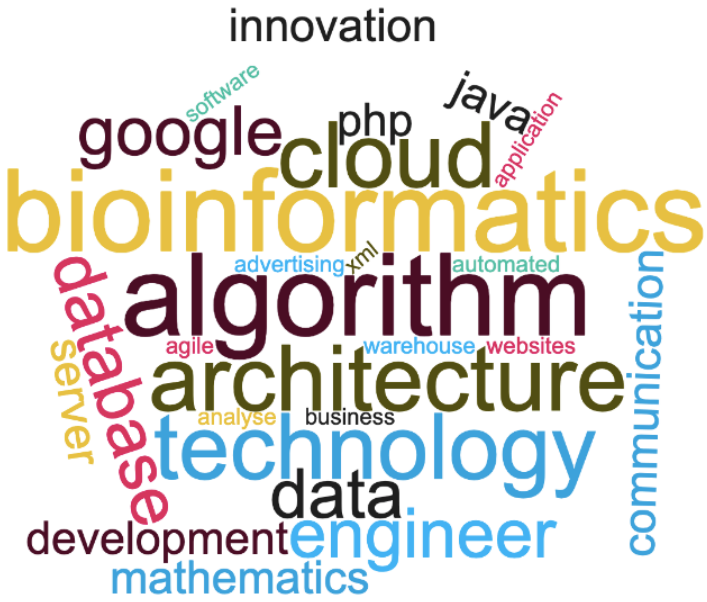


Figure 108. Words of most similar cluster

Figure 108 shows the cluster of words belonging to cluster five, which is most similar to all syllabuses of the three universities. As we have shown in the graphs above, for all three universities maximum similarity was achieved with cluster five, and this has interested us to analyze what the words contained in this cluster are.

Of course, each of the words that we have shown in figure 108 have their importance and their weight based on the frequency that those words have in the labor market demand corpus. We will also later present the words in their order of importance, to see what words are affecting the similarity. Of course, according to our analysis, South East European University ranks as the first university of the three universities in the region for the similarity of textual content between the labor market requirements and curricula offered by the university. According to our model, the University of Pristina ranks as the second university in terms of the textual content of the labor market requirements and curricula offered by this university. And the university that ranks third is University of Tirana. The following is a list of cluster 1 words, since all three universities have less similarity to this cluster. As the textual similarity is smaller for all universities, this cluster contains fewer words in syllabuses offered by universities.

5.3.5. Less similar cluster with syllabuses

Figure 109. Words of less similar cluster

Figure 109 shows the list of cluster 1 words with which the three universities have less similarities. As we can see in Figure 109, we have words sorted by weight, and these are exactly the words that make syllabus resemblance smaller. Below we will compare syllabuses with each other for similarity.

5.3.6. Cluster matching comparison

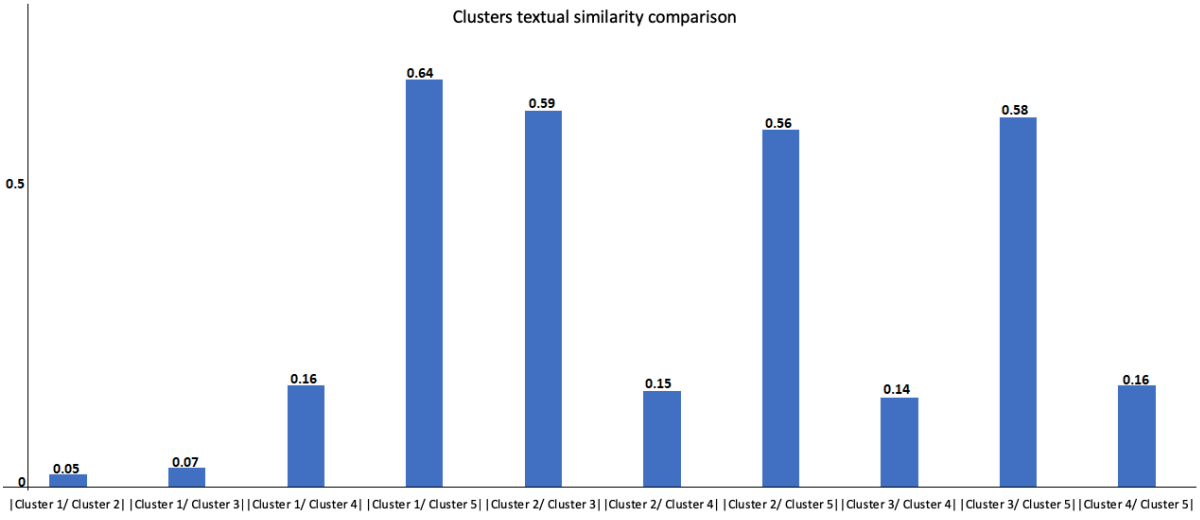


Figure 110. Cluster textual similarity comparison

Figure 110 presents an analysis of the textual similarity between the clusters created by our system. As can be seen in the graph above, the similarity between cluster 1 and cluster 2 is very low since we have a similarity of 0.05. Whereas for the similarity between cluster 1 and cluster 3 we have a textural similarity of 0.07. The greatest similarity begins with other

clusters, as the textual similarity between cluster 1 and cluster 4 reaches 0.16. While the highest textural similarity is between cluster 1 and cluster 5 as it reaches a level of 0.64. Also the similarity between cluster 2 and cluster 3 is high as it reaches a level of 0.59. We have a smaller similarity between cluster 2 and cluster 4, since we have a similarity of 0.15. A very high similarity is observed between cluster 2 and cluster 5, since the value reaches 0.59. Finally we have the similarity between Cluster 3 and Cluster 4 where we have a value of 0.14, while Cluster 3 and Cluster 5 have a similarity of 0.58. As well as the last comparison is between cluster 4 and cluster 5, and the value obtained is lower than the other values because we have a value of 0.16.

In the following we will present the vacancy corpus to see its properties, and to proceed later with in-depth analysis of the corpus.

5.4. Vacancy Corpus

As we know, our market demand corpus has been automatically downloaded from the web sites, and then the text for further analysis has been prepared.

Its initial content is textual, but by means of the algorithms used by our system, this corpus is converted to numerical values which later work for various machine learning analyzes.

The total number of words in our corpus, extracted from the Internet, is 11622 unique words, each word having its own frequency. The highest frequency of words used in our corpus is 3599 for the word "technology", while the smallest frequency for words is 1, for words: experience, auxiliary, etc.

In our case we will present the analysis for the top 5 words that have the highest frequency in the labor market demand corpus.

For these words, the frequency of each will be found, determining the weight of each word, to proceed later for each word and to see how often it is mentioned in the syllabuses of the universities that are part of our research.

The top five words that will be part of our analysis are: technology, analyst, information, security, and support. Below we will present the table with the weight of each word.

Table 8. Top five word frequency and weight

Number	Word	Frequency	Weight
1	technology	3599	0.045%
2	analyst	1832	0.022%
3	information	1605	0.020%
4	security	1497	0.018%
5	support	1146	0.014%

In table 8 we present the top five words that are derived from the labor market demand corpus. As we can see, the highest frequency is 3599, and the smallest frequency is 1146. Of course, the presence of these words in the textual content of the university syllabus determines the level of similarity between labor market requirements and university syllabus. It also shows the weight of each word compared to the 11622 words that are part of the labor market requirements. Initially, word stemming will be done, using only the base of words that are part of the corpus, and then we will show the tables of how many of these words appear in the syllabuses of the three universities to see if they support the results presented earlier.

Table 9. Stemmed words frequency in SEEU Computer Science syllabus

Number	Stemmed word	Frequency in syllabus
1	“technolog”	28
2	“analys”	21
3	“inform”	28
4	“secur”	18
5	“support”	7

In table 9 we have presented the list of words that were initially stemmed, then checked for their frequency in the Computer Science syllabus of South East European University. As we can see in the table above, the word "technologist" in the syllabus is mentioned 28 times, the word "analys" in the syllabus is mentioned 21 times, the word "information" is mentioned 28 times, the word "each" is mentioned 18 times, and the word "support" Mentioned 7 times. It is precisely these words that influence the greater resemblance to South East European

University than to other universities. In the following we will present the table for the University of Pristina.

Table 10. Stemmed words frequency in UP Computer Engineering syllabus

Number	Stemmed word	Frequency in syllabus
1	“technolog”	16
2	“analys”	20
3	“inform”	19
4	“secur”	27
5	“support”	3

Table 10 lists the top five words mentioned in the Computer Engineering syllabus of the University of Pristina. As can be seen in the table above, the word "technologist" in the syllabus is mentioned 16 times, the word "analys" in the syllabus is mentioned 20 times, the word "information" is mentioned 19 times, the word "each" is mentioned 27 times, and the word "support" is mentioned 3 times. As we can see, at the University of Pristina, these words are less frequently mentioned, which directly affects the result of the similarity we obtain for the University of Pristina. Following is the table for the University of Tirana.

Table 11. Stemmed words frequency in UT Informatics syllabus

Number	Stemmed word	Frequency in syllabus
1	“technolog”	0
2	“analys”	2
3	“inform”	5
4	“secur”	0
5	“support”	0

Table 11 shows the frequency of the top five words in the University of Tirana syllabus. As can be seen, we have a very low frequency compared to South East European University and University of Pristina. There are only two words that are mentioned in this syllabus, while the other words that have a great weight on the labor market requirements corpus are not mentioned at all and have a frequency of 0.

Having done all the necessary and foreseen analyzes regarding the textual similarity between the labor market requirements and the curricula offered by the university, we will present a final analysis through the graph, achieving the results that concretize it. we gain from our system.

5.5. Rounded results of our model for all universities

Since we have all the results from all the analyzes, we will now summarize the results in order to give concrete results from our model as to the textual similarity between labor market requirements and curricula offered by universities. Below, we will present the table with the final results of the comparison of labor market requirements and university curricula.

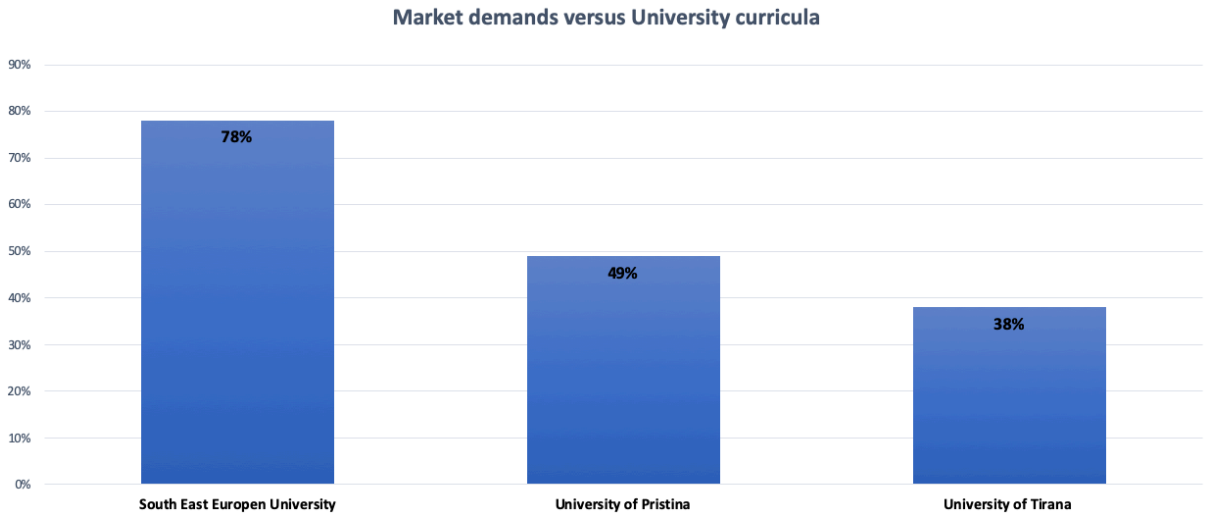


Figure 111. Rounded results between market demands and university curricula

Figure 111 presents the final results that have been concretized and rounded up to be presented as the final data provided by our model. According to the data presented by our model, and after the results were concretized, South East European University has a general syllabic similarity of 78%. University of Pristina has a general syllabus-like similarity to the labor market requirements of 49%. As well as the University of Tirana there is a general similarity of syllabuses with the labor market requirements of 38%. From this we can conclude that our model ranks South East European University as the first university in terms of textual similarity between the labor market requirements and the syllabuses offered by this university. The second university ranked by our model for the textual similarity of syllabuses and labor market requirements is the University of Pristina. While the University with the

least similarity of textual content between syllabuses and labor market requirements is ranked University of Tirana.

6. Conclusion remarks

Creating a model that will make an automated comparison between labor market requirements and curricula offered by universities is of great importance, especially at the present time when labor market demands are on the rise. As we have outlined at the beginning of our topic, the demands of the labor market in the technology field are increasing day by day, and by 2020 there will be approximately 8 million technology jobs.

In our subject we have initially done research and literature review on the latest achievements in this field, and have noticed that only manual analysis has been done so far.

After the literature review was done, we distributed questionnaires to public universities in Kosovo who responded to academic questions and student preparation.

The questions that have been circulated to the academic staff have been about the preparation of students which they have when they come to university and the preparation with which they leave university.

Also part of the question has been about the involvement of private companies in university groups to create curricula offered by universities. In this part the staff responded to the possibility of improving the curricula by answering what are the areas that students face difficulties in studying at the university.

Through these questionnaires we have also sought to obtain information on whether students are able to practice during their studies in order to practice the knowledge they obtain.

Finally, academic staff were asked about the model issue which makes an automated comparison of labor market requirements and curricula offered by universities, with most of them presenting positive thinking and full readiness to contribute in this regard.

In addition to the academic staff, the questionnaires are also distributed to private companies in order to find the gaps that students who have graduated from university have.

Through questionnaires that are distributed to companies we have tried to obtain answers about the scientific degree that employees have. Then, based on the degree students have,

we have tried to identify the gaps they have, and in this way we have come to conclusions that help improve the university curricula.

Knowing that the information was extracted from websites, this has also been a question that has been circulated to private universities and companies.

Almost 90% of responses have been that they publish competitions on websites and that information about labor market requirements is obtained from websites. Through this information, we are based on extracting these labor market requirements from automated web sites to continue with in-depth analysis.

Private companies have also been asked about the practical training that their employees have and the areas that they find difficult after hiring technology companies.

Part of the questions that have been submitted to private companies has also been about the training that private companies provide for their employees. According to the respondents, almost all companies responded that they do not offer a training program for their employees, but are obliged to send them to training abroad.

What are the demands of the labor market in the field of technology, companies have stated that they are unaware of this issue, and certainly our topic has contributed to this right as it has also emerged statistics on labor market requirements.

These questionnaires have been part of our research motivation in this regard, as even private companies have stated that they are willing to contribute to the implementation of a model that makes automated comparisons between labor market requirements and curricula. university.

Following the implementation of our model we have made some specific analyzes which are accurate data as to the adaptation of labor market requirements and university curricula.

Part of our research has been three universities in the region, South East European University, University of Pristina and University of Tirana.

Initially we compared each syllabus of the university with the demands of the labor market and saw how they fit with each other.

Following this analysis, we have compared specific technology positions with specific syllabuses offered by universities, and we have seen how these are tailored to each other.

The next analysis that has been done has been to compare the syllabuses of the region's universities with the syllabuses offered by the top world universities.

Following these analyzes, we conducted a corpus of labor market demands by removing stop words, and compared the new results with the preliminary results.

We also removed some competitions, and added some competitions to test the consistency of our model.

The robustness of our model is also verified when compared to another existing model which compares the textual content but distinguishes the algorithm that our model contains. Of course, the results offered by our model are far more accurate and consistent than the other existing model.

Creating clusters has been indispensable for our theme, but we have previously analyzed the optimal number of clusters that our theme should contain. Based on the analysis done through the silhouette score, we have been able to find the optimal number of clusters to later perform sequential analyzes of the fit of the clusters to the syllabuses provided by the universities.

These clusters have also been compared to each other in terms of the textual similarity they have with each other.

Finally, the results that our model offers about adapting to the labor market requirements and university syllabus are concretized.

And finally, our model provided accurate data which ranked South East European University as the first university in terms of the textual similarity of the labor market and syllabus requirements offered by this university. The University of Pristina is ranked second in terms of textual similarity, and the University of Tirana is ranked third in terms of textural similarity between the labor market requirements and curricula offered by this university.

So in the end we can conclude that our model will be of great help in adapting the curricula to the demands of the labor market. We can also conclude that the results provided by our model are highly accurate given that they are provided by modern automated methods.

In the following we will present the future work of our topic, allowing other researchers to contribute in this direction.

6.1. Future work

As with other research, we too will have room for future researchers to contribute. Of course the spaces will not be large but some of them we will mention.

Initially as future work we are thinking of automatically downloading labor market data as soon as a vacancy is published on that website. Of course, this data will not change much in the results that our model offers, but after a while it will be an advantage if these data are automatically downloaded when published on the website.

Second, as future work, we are thinking of generating reports in real time where all domains ranging from universities, companies to students will be able to track these results based on the results that change in real time.

Third, as future work, we think of new job prospects and new study programs based on data downloaded from websites. Labor market requirements will be used to recommend the creation of new programs, and the syllabus offered by the university will be used to foresee any new jobs that may open in the near future.

Finally, as a future work, we also see the conversion of our data and their preparation for the semantic web, so that we can map the words that are part of our corpus.

References

[1] M. Agaoglu, “Predicting Instructor Performance Using Data Mining Techniques in Higher Education”. *In IEEE Access*, May 2016.

[2] T. Xie, Q. Zheng, W. Zhang, H, Qu, “Modeling and Predicting the Active Video – Viewing Time in a Large – Scale E – Learning System”.*In IEEE Access*, June 2017.

[3] A. M. Njeru, M. S. Omar, S. Yi, “IoTs for Capturing and Mastering Massive Data Online Learning Courses*”. In IEEE Computer Society, ICIS*, Wuhan, China, May 2017.

[4] R. Heartfield, G. Loukas, D. Gan, “You are probably not the weakest link: Towards Practical Prediction of Susceptibility to Semantic Social Engineering Attacks*”. In IEEE Access,* October 2016.

[5] E. J. Fortuny, D. Martens, “Active Learning – Based Pedagogical Rule Extraction”. *In IEEE Transaction on Neural Network and Learning Systems*, Vol. 26, No. 11, November 2015.

[6] A. Mukhopadhyay, S. Bandyopadhyay, “A Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part I”. *In IEEE Transaction on Evolutionary Computation*, Vol. 18, No. 1, February 2014.

[7] Zh. Song, A. Kusiak, “Optimization of Temporal Processess: A Model Predictive Control Approach*”. In IEEE Transaction on Evolutionary Computation*, Vol. 13, No. 1, February 2009.

[8] S. Malgaonkar, S. Soral, Sh. Sumeet, T. Parekhji, “Study on Big Data Analytics Research Domain”. In *International Conference on Reliability, Infocom Technologies and Optimization ICRITO,*Noida, India, September 2016.

[9]K. P. Anicic, B. Divjak, K. Arbanas, “Preparint ICT Graduates for Real – World Challenges: Results of a Meta – Analysis”. In *IEEE Transactions on Education,* Vol 60, No. 3, August 2017.

[10] A. Haskova, D. V. Merode, “Professional Training in Embedded Systems and its Promotion”. In *IEEE Transacions on Education,* 2016.

[11] S.C. Smith, W. K. Al-Assadi, J. Di, “Integrating Asynchronous Digital Design into the Computer Engineering Curriculum”. In *IEEE Transactions on Education*, Vol. 53, No. 3, August 2010.

[12] M. D. Koretsky, D. Amatore, C. Barnes, Sh. Kimura, “Enhancement of Student Learning in Experimental Design Using a Virtual Laboratory”. In *IEEE Transactions on Education*, Vol. 51, No. 1, February 2008.

[13] B. G. Member, V. S. Sheng, K. Y. Tay, W. Romano, Sh. Li, “Incremental Support Vector Learning for Ordinal Regression”. In *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 26, No. 7, July 2015.

[14] J. Li, T. Zhang, W. Luo, J. Yang, X. T. Yuan, J. Zhang, “Sparseness Analysis in the Pretraining of Deep Neural Networks”. In *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 28, No. 6, June 2017.

[15] Y. Qian, F. Li, J. Liang, B. Liu, Ch. Dang, “Space Structure and Clustering of Categorical Data”. In *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 27, No. 10, October 2016.

[16] Y. Xiao, B. Liu, Zh. Hao, “A Maximum Margin Approach for Semisupervised Ordinal Regression Clustering”. In *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 27, No. 5, May 2016.

[17] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, Sh. Li, “Incremental Support Vector Learning for Ordinal Regression”. In IEEE *Transactions on Neural Networks and Learning Systems*, Vol. 26, No. 7, July 2015.

[18] P. Navrat, L. Molnar, “Curricula Transformation in the Countries in Transition: An Experience from Slovakia”. In IEEE *Transactions on Education*, Vol. 41, No. 2, May 1998.

[19] S. Nalintippayawong, K. Atchariyachanvanich, “IT Management Status in Public Higher Education Institutions in Thailand”. In IEEE *ICIS 2016*, June 26-29, 2016, Okayama, Japan.

[20] J. I. Godino – Llorente, R. Fraile, J. C. Gonzales de Sante, V. Osma – Ruiz, N. Saenz – Lechon, ““Design for All in the Context of the Information Society”: Integration of a Specialist Course in a Generalist M.Sc. Program in Electrical and Electronics Engineering”. In *IEEE Transactions on Education*, Vol. 55, No. 1, February 2012.

[21] M. Dolores Cano, “Students’ Involvement in Continuous Assessment Methodologies: A Case Study for a Distributed Information Systems Course”. In IEEE *Transactions on Education*, Vol. 54, No. 3, August 2011.

[22] Y. He, Ch. Wang, Ch. Jiang, “Mining Coherent Topics with Pre-Learned Interest Knowledge in Twitter”. In *IEEE Access*, June 2017.

[23] H. Pirkkalainen, J. P. P. Jokinen, J. M. Pawlowski, “Understanding Social OER Environments—A Quantitative Study on Factors Influencing the Motivation to Share and Collaborate”. In IEEE *Transactions on Learning Technologies*, Vol. 7, No. 4, October-December 2014.

[24] G. Goth, “Network-Enabled Compulsory Education Getting Big Push”. In *IEEE Computer Society*, February 2009.

[25] R. Mehmood, F. Alam, N. N. Albogami, I. Katib, A. Albeshri, S. M. Altowaijri, “UTiLearn: A Personalised Ubiquitous Teaching and Learning System for Smart Societies”. In *IEEE Access,* February 2017.

[26] J. J. Guerrero, L. A. Guerrero, “A Virtual Repository of Learning Objects to Support Literacy of SEN Children”. In *IEEE Revista Iberoamericana De Tecnologias Del Aprendizaje*, Vol. 10, No. 3, August 2015.

[27] A. A. Choudhury, J. Rodriguez, “A New Curriculum in Fluid Mechanics for the Millennial Generation”. In *IEEE Revista Iberoamericana De Tecnologias Del Aprendizaje,* Vol. 12, No. 1, February 2017.

[28] A. Sethi, “Factors Responsible for Mismatch between Demand and Supply of Requisite Skill in India”. In *IJARIIE*-ISSN (O)-2395-4396, Vol. 3, Issue 3, 2017.

[29] L. Anastasiu, A. Anastasiu, M. Dumitran, C. Crizboi, A. Homaghi, M. N. Roman, “How to Align the University Curricula with the Market Demands by Developing Employability Skills in the Civil Engineering Sector”. In *Education Sciences*, doi: 10.3390/educsci7030074, July 2017.

[30] K. P. Anicic, B. Divjak, K. Arbanas, “Preparing ICT Graduates for Real – World Challenges: Results of Meta – Analysis”. In *IEEE TRANSACTIONS ON EDUCATION,* Vol. 60, No. 3, August 2017.

[31] M. T. R. A. Aziz, Y. Yusof “Graduates Employment Classification using Data Mining Approach”. In *Proceedings of the International Conference on Applied Science and Technology*, ICAST, 2016.

[32] S. Sahu, M. Bhatt, “Big Data Classification of Student Result Prediction”. *In International Journal of Research In Science & Engineering*, Volume: 3 Issue: 2 March-April 2017.

[33] V. Bharanipriya, V. Kamakshi Prasad, “Web Content Mining Tools: A Comparative Study”. In *International Journal of Information Technology and Knowledge Management,* Volume 4, No. 1, pp. 211-215, 2011.

[34] P. Thakar, A. Mehta, Manisha, “Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue”. In *International Journal of Computer Applications (0975 – 8887)*, Volume 110 – No. 15, January 2015.

[35] G. Grasso, T. Furche, Ch. Schallhart, “Effective Web Scraping with OXPath”. In *WWW 2013 Companion, Rio de Janeiro, Brazil. ACM 978-1-4503-2038-2/13/05, 2013.*

[36] M. Thelwall, “A Web Crawler Design for Data Mining”. In *Journal of Information Science,* pp. 319–325, 2001.

[37] T. V. Adapure, R. D. Kale, R. C. Dharmik, “Study of Web Crawler and its Different Types”. In *IOSR Journal of Computer Engineering*, Volume 16, Issue 1, PP 01-05, 2014.

[38] F. Ahmad, N. H. Ismail, A. A. Aziz, “Using Classification Data Mining Techniques”. In *Applied Mathematical Sciences,* Vol. 9, pp. 6415 - 6426, no. 129, 2015.

[39] D. Garcia-Saiz, M. Zorilla, “Comparing Classification Methods for Predicting Distance Students Performance”. In *JMLR: Workshop and Conference Proceedings 17,* pp. 26-32, 2011.

[40] N. R. Haddaway, “The Use of Web-scraping Software in Searching for Grey Literature”. In *The Grey Journal,* Volume 11, 2015.

[41] T. Furche, G. Gottlob, G. Grasso, Ch. Schallhart, A. Sellers, “OXPath: A language for scalable data extraction, automation, and crawling on the deep web”. In *The VLDB Journal*, 2012.

[42] L. Auria, R. A. Moro, “Support Vector Machines (SVM) as a Technique for Solvency Analysis”. In *German Institute for Economic Research,* 2008.

[43] M. Awad, L. Khan, F. Bastani, “An Effective Support Vector (SVM) Performance Using Hierarchical Clustering”. In *IEEE 24th International Conference on Tools with Artificial Intelligence,* 2004.

[44] A Brief Introduction to Support Vector Machine (SVM). January 25, 2011.

[45] ACM Recommendations for Computer Science Curricula, Volume I, 1983.

[46] Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering
A Volume of the Computing Curricula Series, IEEE, 2014.

[47] Th. Iliou, Ch. N. Anagnostopoulos, M. Nerantzaki, “A Novel Machine Learning Data Preprocessing Method for Enhancing Classification Algorithms Performance”. In *16th EANN workshops*, ACM, Rhodes Island, Greece, 2015.

[48] M. M. Yusof, R. Mohamed, N. Wahid, “Benchmark of Feature Selection Techniques with Machine Learning Algorithms for Cancer Datasets”. In *ICAIR and CACRE '16*, ACM, Kitakyushu, Japan, 2016.

[49] D. Brandon, “TEACHING DATA ANALYTICS ACROSS THE COMPUTING CURRICULA *”. In *CCSC: Mid-South Conference*, 2015.

[50] H. Hu, J. Li, A. Plank, H. Wang, G. Daggard, “A Comparative Study of Classification Methods for Microarray Data Analysis”. In *Proc. Fifth Australasian Data Mining Conference*, 2006.

[51] H. Liu, X. Yin, J. Han, “An Efficient Multi-relational Naïve Bayesian Classifier Based on Semantic Relationship Graph”. In *MRDM’05,* ACM, Chicago, Illinois, USA, 2005.

[52] M. HooshSadat, H. W. Samuel, S. Patel, “Fastest Association Rule Mining Algorithm Predictor (FARM-AP)”. In *C3S2E 11,* Montreal, QC, Canada, 2011.

[53] H. Hu, J. Li, “Using Association Rules to Make Rule-based Classifiers Robust”. In *Australian Computer Society, Inc.,* ACM, 2005.

[54] A. Sun, E. Lim, W. Ng, “Web Classification Using Support Vector Machine∗”. In *WIDM’02,* ACM, Virginia, USA, 2002.

[55] L. Borges, V. Marques, J. Bernardino, “Comparison of Data Mining techniques and tools for data classification”. In *C3S2E-13*, ACM, Porto, Portugal, 2013.

[56] Y. N. Silva, S. W. Dietrich, J. M. Reed, “Integrating Big Data into the Computing Curricula”. In *SIGCSE’14,* ACM, Atlanta, GA, USA, 2014.

[57] E. Trandafili, A. Allkoci, Elinda Kajo, A. Xhuvani, “Discovery and Evaluation of Student’s Profiles with Machine Learning”. In *BCI’12, ACM 978-1-4503-1240-0/12/09*, Novi Sad, Serbia, 2012.

[58] B. Edwards, M. Zatorsky, R. Nayak, “Clustering and Classification of Maintenance Logs using Text Data Mining”. In *Proc. 7th Australasian Data Mining Conference (AusDM'08),* Glenelg, South Australia, 2008.

[59] L. Merschmann, A. Plastino, “A Bayesian Approach for Protein Classification”. In *SAC’06, ACM 1-59593-108-2/06/0004,* Dijon, France, 2006.

[60] A. Veloso, W. Meira Jr., M. Cristo, M. Goncalves, M. Zaki, “Multi-Evidence, Multi-Criteria, Lazy Associative Document Classification”. In *CIKM’06, ACM 1-59593-433-2/06/0011,* Virginia, USA, 2006.

[61] R. Frank, M. Ester, A. Knobbe, “A Multi-Relational Approach to Spatial Classification”. In *KDD’09, 978-1-60558-495-9/09/06,* Paris, France, 2009.

[62] M. Ericsson, A. Wingkvist, “Mining Job Ads to Find What Skills are Sought After from an Employers’ Perspective on IT Graduates”. In *ITiCSE’14, ACM 978-1-4503-2833-3/14/06,* Uppsala, Sweden, 2014.

[63] Q. Ding, Q. Ding, W. Perrizo, “Decision Tree Classification of Spatial Data Streams Using Peano Count Trees”. In *SAC 2002, ACM 1-58113-445-2/02/03,* Madrid Spain, 2002.

[64] Ch. C. Aggarwal, “The Setwise Stream Classification Problem”. In *KDD’14, ACM 978-1-4503-2956-9/14/08,* New York, NY, USA, 2014.

[65] Ch. C. Aggarwal, J. Han, Ph. S. Yu, “On Demand Classification of Data Streams”. In *KDD’04, ACM 1-58113-888-1/04/0008,* Seattle, Washington, USA, 2004.

[66] Ch. C. Aggarwal, “Towards Exploratory Test Instance Specific Algorithms for High Dimensional Classification”. In *KDD’05, ACM 1-59593-135-X/05/0008,* Chicago, Illinois, USA, 2005.

[67] J. Li, R. Topor, H. Shen, “Construct robust rule sets for classification”. In *SIGKDD’02, ACM 1-58813-567-X/02/0007,* Alberta Canada, 2002.

[68] N. Jin, C. Young, W. Wang, “GAIA: Graph Classification Using Evolutionary Computation”. In *SIGMOD ’10, ACM 978-1-4503-0032-2/10/06,* Indiana, USA, 2010.

[69] H. Fei, J. Huan, “Structure Feature Selection for Graph Classification”. In *CIKM ’08, ACM 978-1-59593-991-3/08/10,* California, USA, 2008.

[70] A. Wegmann, “Theory and Practice behind the Course Designing Enterprisewide IT Systems”. In *IEEE Transactions on Education*, Vol. 47, No. 4, November 2004.

[71] J. E. Froyd, Ph, C. Wankat, K. A. Smith, “Five Major Shifts in 100 Years of Engineering Education”. In *Proceedings of the IEEE, 0018-9219*, Vol. 100, May 13th, 2012.

[72] Z. Shiller, “A Bottom-Up Approach to Teaching Robotics and Mechatronics to Mechanical Engineers”. In *IEEE Transactions on Education*, Vol. 56, No. 1, February 2013.

[73] M. Marques, M. C. Viegas, M. C. Costa – Lobo, A. V. Fidalgo, G. R. Alves, J. S. Rocha, I. Gustavsson. In IEEE *Transactions on Education*, Vol. 57, No. 3, August 2014.

[74] A. K. Kakar, “Teaching Theories underlying Agile Methods in a Systems Development Course”. In *47th Hawaii International Conference on System Science, IEEE, 978-1-4799-2504-9/14,* 2014.

[75] W. He, J. T. Kwok, J. Zhu, Y. Liu, “A Note on the Unification of Adaptive Online Learning”. In IEEE *Transactions on Neural Networks and Learning Systems*, Vol. 28, No. 5, May 2017.
[77] D. Klosters, “Matching Skills and Labor Market Needs”. In *World Economic Forum,* January 2014.

[76] A. Ghani Kanesan Bin Abdullah, “Bridging the Gap between Industry and Higher Education
Demands on Electronic Graduates’ Competencies”. In *IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)*, Volume 8, Issue 1, 2013.

[77] S. Melnik, H. Garcia-Molina, E. Rahm, “Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching”. *In IEEE*, San Jose, CA, USA, 2002.

[78] Mayra Z. Rodriguez, “ Clustering algorithms: A comparative approach”. doi: 10.1371/journal.pone.0210236

[79] J. Hahn, M. Kamber, J. Pei, “Data Mining Concepts and Techniques”, Third edition. In *Elsevier*, Waltham, USA, 2012.