Ionela MANIU
Department of Mathematics and Informatics, Research Center in Informatics and Information Technology, Faculty of Sciences, Lucian Blaga University of Sibiu, Romania
Emilia-Loredana POP
Department of Computer Science Faculty of Mathematics and Computer Science, Babes Bolyai University Cluj-Napoca, Romania
Augusta RATIU
Department of Mathematics and Informatics Faculty of Sciences, Lucian Blaga University of Sibiu, Romania
Eduard Traian STEFANESCU
Faculty of Sciences, Lucian Blaga University of Sibiu, Romania

INSIGHTS FROM IT JOBS MARKET WITH TEXT MINING APPROACH

Case Study

Keywords
Text Mining;
IT Jobs;
Job Description;
Word Cloud;

Abstract

On the labor market, IT Jobs represent one of the most important domains. This paper analyzed the IT Jobs from a large collection of job listings from a Romanian website. The Web Crawling techniques were used to extract the data from the website, the Text Mining, World Cloud and statistic techniques to analyze and present the results. Insights that are required for an IT Job were extracted. The results highlight the following: the most required IT Job Type was the Full Time one, the Career Level was the Mid Level and the Study Level was the Graduated one. The Text Mining Approach revealed that the most frequent words for the IT Jobs offers were: team, work, development, project, experience, environment, service, skill, customer, knowledge and software. A comparison between terms in IT Jobs in English language versus IT Jobs in Romanian language was also performed. As conclusion, in this paper, by combining different techniques to extract and analyze the textual data set of vacancy offers knowledge from the IT Jobs domain was determined and discovered. The extracted information presented insights in current skills, personal needs, career paths of the Romanian current IT labor market needs. The results of this research could be valuable information for public bodies, employers, higher education policy makers, researchers, students and parents.

INTRODUCTION

Nowadays IT Jobs play an important role in labor market; the landscape is moving very quickly and the students are looking for more and more jobs in the IT domain. The possibilities are open for every person and the chance to get the better job is increasing every day by having a lot of opportunities.

Previous researches from this area explored the job listings through various perspectives, starting with the characterization of the job title or with the job description (Chopra and Golab, 2018; Espinoza, Guerrero and Agudo, 2015; Kino, Kuroki, Machida, Furuya and Takano, 2017; Kobayashi, Mol, Berkers, Kismihok and Hartog, 2018; Maceli, 2015; Terek, Mitic, Cvetkoska, Vulkonjanski and Nikolic, 2018). The approach that is considered in this paper brings a new perspective on ongoing and point to a highly necessary work.

Labor market is in continuous changing and in the context of economic factors and not only, its concepts and practical issues are dynamic. Job offer requirements are changing over time and should be analyzed in their competitive contexts.

The aim of this paper is to analyze the job offers from a Romanian website. The job listings considered in this paper are the ones from the public website https://www.ejobs.ro/, from which all the IT Jobs available from the period April - May 2020, from all the locations and with all the specifications were considered.

Thus, this study aims to answer the following two research questions:

- RQ1: What is the benefit of analysing the profile of IT Jobs? and
- RQ2: Does the text mining methodology increase the value by extracting important terms from IT Job Description to have a better understanding of the labor market?

The study makes the following contributions: uses a combination of different methods to extract data and information relative to the IT job postings from an online website and offered insights on the IT Jobs labor market from Romania as a case study.

The study is structured in five sections: Introduction, Related Work, Methodology, Experiments and Results, Conclusions. In Introduction is presented the research direction regarding the labor market in the IT field. Section Related Work brings some case studies, especially for IT Jobs, for which were applied Text Mining, World Cloud techniques and not only. In the Methodology section are revealed information about the techniques, like, Text Mining, World Cloud, that were used in analyzing the data and how these ones were applied. In the section called Experiments and Results, was presented the considered data set which was analyzed along with the obtained results and discussions. The results refer to the IT Job City, IT Job Type and IT Job Description and are followed by the conclusions.

RELATED WORK

The analysis of the job title and job description was done by different authors and from different points of view.

A simple and effective method used to monitor the skills need it by potential employers in their job advertisements, was introduced in (Wowczko, 2015). The RapidMiner and R open source tools were used for the online vacancy data to extract information about the skills. For a collection of vacancies the authors analyzed the knowledge, by using techniques like classification with kNN and also the extraction of information from the dataset. The approach used was able to identify the occupations and labor market demands within a given dataset. By taking into account the job description, the advertisements were splitted into groups. For these groups, were analyzed the top 20 corresponding bigrams with R’s wordcloud package and the conclusions followed clearly.

The objective in (Espinoza et al., 2015) was to identify the most required specialization profiles for the companies and organizations in Lima, for the Blog ”Estad´ısticos de Peru”, where 2809 job postings were analyzed using the Text Mining techniques. The authors extracted relevant information and generic skills for the Peruvian statisticians. The job postings were separated in five segments by using the Singular Value Decomposition (SVD) and analyzed each of them with SAS Enterprise Miner. The authors identified important requirements, competencies and demands of the companies that were included in the job postings. The text mining techniques were used to detect the professional profiles from the job postings. The comparison between the requirements and skills from the dataset of the periods 2009-2011 and 2012-2014, gave recommendations to the agents involved in the job market. The R libraries and SAS Text Miner for analyzing the dataset with postings, were used in obtaining the results: ”experience” was one of the first things required of a statistic, the keyword ”database” was found frequently, meaning that the SQL language has become very important in Lima. Also, the professional profiles, like, Business Intelligence Professionals, Risk managers, Students or graduates in trainee programs, Market researchers, were obtained by using Word Clouds.

In (Kino, et al., 2017), the authors wanted to improve the matching process for staffing agencies in Japan, by analyzing the data referring to commute time, job location, job type, hourly rates, skill set of candidate, and so on. The main idea was to get an appropriate placement for both companies and candidates, by matching the connection between them. The Japanese text analysis tool used for the free text was KH Coder. Job Matching worked to connect a company and a candidate: matching a candidate to a specific job and also matching a job for a specific candidate. Three types of datasets were considered: requirements and job information from companies recruiting employees, requirements of and skill data of candidates, and historical matching results from the recruiter. The important keywords, giving positive or negative influences, for job matching were found with the KH Coder tool, combined with the Japanese morphological analysis, hierarchical clustered analysis and co-occurrence network. A frequency list of keywords was created. The keywords most suggestive were referring to administrative notifications (like, preferred job condition, current working status, skills, post-placement evaluation). The co-occurrence network diagrams were created to explain better the conclusions. By the text mining analysis, were found keywords for positive aspects (Continue, Regular, Exercise, Kindergarten, Parents home), for negative aspects (Necessity, Bad, Hard) and also, for both of them (Introduction, Communication). The keywords found could be useful to progress the job matching system by applying them to a machine learning process.

Another text mining methodology was used on co-operative education (co-op) programs job posting corpus mentioned by the Word Association for Cooperative and Work-integrated Education from a university from North America. The authors worked with an unstructured job description, from which were extracted and clustered the informative terms (Chopra and Golab, 2018). There were analyzed nearly 30000 co-op job postings. A parser was used to extract the relevant attributes from the unstructured job description. For each attribute found in job title and job description the frequency was identified. By employing the Latent Semantic Analysis and the k-means clustering for the attributes extracted, was obtained a characterization for each type of available job. The analysis was done for the year 2014, when the IT jobs were focused on software, especially on Java and web development. A comparison with the year 2004 brought differences related to new technologies, soft skills, mindset and company culture. Moreover, the authors did also an analysis relative to the others disciplines.

For job listings in Library and Information Science was done a research related to the Text Analysis in the idea of highlighting the skills and the group of skills for the employers from North America (Maceli, 2015). The job listings were taken from Code4lib jobs website and the analysis was done with the R statistical package. A Job Title Analysis was done and also a Job Description analysis, in which the most relevant skills were highlighted.

In Text Mining research, were considered some important steps that were analyzed in (Kobayashi et al., 2018): Text Prepossessing, Text Mining Operations and Postprocessing. The authors considered a Job Analysis with Text Mining Applied for some job vacancies and focused on Job Description field. The conclusions arisen due to the steps considered previously. For a large volume of text data, Text Mining has been proved to be efficient in reducing the personnel and the cost constraint. By the practical recommendations came tips on the start process on Text Mining and on the tools that should be used.

METHODOLOGY

Based on the literature, there are a lot of methods to be used for Text Analysis. In this paper the Job Description was analyzed in order to extract important insights need it when a student or a person is looking to get hired in an IT Job. Text Mining techniques related to the analysis with the Word Cloud were used.

1) Data Extraction: In order to extract the information from the website https://www.ejobs.ro/, a custom web crawler was created by using the .NET Core framework, the HtmlAgilityPack (HAP) and also the CSVHelper libraries (HAP - Html Agility Pack, 2020). This custom web crawler took each IT and Telecom page and each job posted on that page, until the end of the results to extract the needed data. The HtmlAgilityPack was used to extract the data from the HTML elements by using their XPath queries (for example, //*[@id=”contentstudii”]/ul[1]/li/a). The CSVHelper library was used to save the results in a Comma Separated Value file.

2) Data Analysis: Text Mining is a part of Data Mining that allows the information extraction from a collection of documents by using specialized analysis tools. This process is passing through multiple stages, discovering the knowledge for the texts that are analyzed (Berry, 2003; Han, Kamber and Pei, 2012; Witten, Frank and Hall, 2011). The information is stored as input text and the Text Mining process is discovering the patterns and the trends, derives the patterns within the structured data, evaluate it and then the output has to be interpreted. Text Mining includes the text characterization, text clustering, entity extractions, document summary, entity relation modeling, production of granular taxonomies, sentiment analysis. In this case, the Text Mining techniques were used to analyze the Job Description of the public IT Job website. There were highlighted the most important and frequently used keywords.

Word Cloud is functioning as a communication tool; it is simply, clear, understandable and returns the keywords that are most representative from the text considered. Moreover, by creating and using a Word Cloud a visual representation of the data from the text was presented. R open source tool for data analysis and data visualization was used (Fellows, 2018; Text mining and word cloud fundamentals in R, 2020). An important and time consuming step was the one focused on the improvement of the pre-processing and of the techniques used in extracting the information from the Job Description and returning the most significant keywords.

EXPERIMENTS AND RESULTS

Data Set Exploration

In this paper, all the IT Jobs available on the website https://www.ejobs.ro/ on the period April - May 2020 were considered. There were chosen only two months for analyze because the website allows the keep of the IT Jobs for a short period of time. A number of 516 IT Jobs were found. For every IT Job Title was analyzed the Job Description to extract the important and useful information in the process of finding jobs from all the cities available on the website (the jobs from Romania, the remote ones and the ones that are across the borders). The IT Jobs considered are some of them in English and others in Romanian. The IT Jobs were analyzed relative to their type, city, career level and study level.

The extraction that has been accomplished with the help of a web crawling tool allowed putting the data into an Excel file and then analyzing it with the help of Text Mining techniques. For each IT Job, were extracted the job title, the posted date, the company, the city, the type of the job, the study level, the career level, the department, the industry, the salary, the job description and the ideal candidate. Also, attention to each job was needed because some of them were in English language and others in Romanian language (see, Table 1).

The number of IT Jobs in May 2020 was three times higher than the number of IT Jobs in April 2020 (see, Table 2).

For all the IT Jobs considered descriptive statistics techniques were used to analyze them. The first analyzing process on these IT Jobs has been done for the Job Type, Job City, career Level and Study Level. The category that was considered separately and analyzed with the help of Text Mining and also with the tool Word Cloud is the Job Description, which was analyzed and the obtained results highlight important insights that a candidate should have for an IT Job. An example for an IT Job Description can be found in Figure 1.

Results and Discussions

In what follows were analyze the IT Jobs relative to the Job City, Job Type, career Level, Study Level and their combinations. Also, an analysis of IT Jobs relative to their Job Description taking into account the results given by the Word Cloud was performed.

1) IT Job City Analysis: The analysis of the IT Jobs is done from the perspective of Job City and it is the Romanian interpretation (due to the city names).

By analyzing the IT Job City relative to all the cities involved in the IT Jobs posted, the conclusion was that Bucharest or Bucuresti (in Romanian) is the city with the highest number of available IT Jobs for the period analyzed, and it also can be found in the IT Jobs that are available for multiple cities. We need also to mention that all the cities included here have the native names in Romanian (see, Figure 2).

Next, is presented the general analysis of the number of IT Jobs available relative to the IT Job City. In Figure 3 is presented the number of IT Jobs available in the most important cities from Romania and their percentages for the months considered. It is obviously that the number of IT Jobs posted during the month May 2020 it is much higher than the one of the IT Jobs posted in the month April 2020.

A more detailed analyze for the IT Job City is realized in Figure 4, also including the most relevant cities from Romania in which the IT Jobs are available. Also, in this analysis, attention should be paid to the fact that the same IT Job can be available in multiple cities.

2) IT Job Type Analysis: By analyzing the IT Jobs from the perspective of Job Type, the following categories were retrieved: Full Time, Part Time, Seasonal, Voluntary and Project (for more details, see, Figure 6). For every IT Job considered, the type field can have one or more values (for example, for the IT Job Title ”Java Developer” the Job Type is ”Full Time”, but also ”Part Time” and instead, for the IT Job Title ”Junior PHP Developer” the Job Type is only ”Full Time”, due to the IT Jobs posted).

The percentage of each Job Type can be related to the period for which the IT Jobs were considered and their representation is given in Figure 5 and Figure 6. It is obvious that the prevalent type is the Full Time type, followed by the Part Time type, the Project and Seasonal types and the less required type is the Voluntary type.

In the month May 2020 it is noticed that the number of the IT Jobs Types: Part Time, Seasonal, Voluntary and Project, is increasing, relative to the month April 2020.

3) IT Job Analysis from combined perspectives: Here were analyzed the IT Jobs from a different perspective that involves multiples values.

The perspective considered here is the one in which the analyses is done from the IT Job Type, IT career Level, IT Study Level perspectives and are related with the corresponding months. The values for each IT Job Type, IT career Level and IT Study Level can be multiples (for example, for an IT Job Title it can be more possible values for the IT Job Type, IT career Level or IT Study Level). All the analyses can be found in Figure 6.

IT career Level: Almost 80% of the IT Jobs require Mid Level experience and more than 40% request the Senior Level experience. The IT career Level less wanted was the one with No Experience and there were only a few cases for the Executive Manager level.

IT Study Level: Over 80% of the IT Jobs require Graduated study level and more than 40% require Qualified persons. The IT Jobs for Students are in percentage of 30%.

Definitely, one can generate and analyze other perspectives, too, and have more results.

4) IT Job Description Analysis: After the text analysis, answers to the research questions arise and more details can be provided. For both of the questions, the answers are clear and objective.

A descriptive analyze for the IT Jobs allows identifying the profile from the labor market and insights of its dynamic.

The text mining methodology was very useful in extracting important words from the IT Job Description and helped in giving insights related to the terms and group of terms used, and brought the benefit of a better understanding of the IT Jobs from the labor market.

Next, it can be seen the Word Cloud corresponding to the complete listing of IT Job Description from https://www.ejobs.ro/, one for the words in English (Figure 7) and one for the words in Romanian (Figure 9). These words helped to extract the main requirements asked by the IT Companies.

As it can be seen, in Figure 7, the most relevant and prevalent words from the job description are appearing larger. So, this means that the most used words are the most relevant ones, with a high percentage in the description. One of the most used word is „team", meaning that the new employee should have the ability to work in a team and adapt to the decision taken together. Some other generic and basic insights refers to the capacities and possibilities of working (”work”, ”development”, ”will”), then, to job insights (”project”, ”experience”), and also, to other aspects (”environment”, ”skill”, ”working solution”, ”knowledge”, ”software”, ”service”, ”customer”, ”system”).

Figure 8 depicts more clearly the predominant insights structured by the counting from the Job Description, for the English language, that gives the importance of each aptitude need it.

In Figure 9, the Word Cloud is corresponding to the language Romanian and the words are quite similar. Here the most used word is ”clienti”, meaning that the new employee should take into account all the needs and requirements of the client, for which he or she is working for. Another very important aspect refers to the continuous development (”dezvoltarea”), that should characterize the new employee. Other groups of words (”lucru”, ”companiei”, ”firmei”), (”activitate”, ”solutii”, ”sistem”, ”echipa”, ”servicii”, ”asigura”, ”proiect/proiectare”, ”aplicatii”) highlights the important capacities, the insights and the available possibilities in work.

Figure 10 illustrates more clearly the predominant insights structured by the counting from the Job Description, for the Romanian language, that give’s the importance of each aptitude need it.

When comparing the conclusions related to both Word Clouds (the one in English and the one in Romanian), the most relevant difference consist in the order of each insights required, or how important is each one for the IT Jobs (see, Figure 11). For example, ”team” is the most important one in English and only on the forth position in Romanian. The ”development” is on the second position in importance for both English and Romanian. The ”work” is on the second position in English and third position in Romanian. The ”project” is on the third position in English and on forth position in Romanian. The ”solution”, ”software”, ”service”, ”system” are all on the forth position in both English and Romanian. The ”customer” is on the forth position in English and first position in Romanian.

CONCLUSIONS

The results of the current research could be valuable information for public bodies, employers, researchers, students and parents. Through continuous data collections and analysis using text, data and web mining technologies, higher education policy makers would be able to identify the trends and skills in programming jobs, and to update programs to work in order to meet the demands of Information Technology labor market. As an example, many job descriptions seem to emphasize teamwork and consequently the HIE (Higher Education Institutions) should incorporate more team and project exercises in their academic curriculum.

The lists of frequent terms presented by this research may be used by employers to re-design job postings; for example, creating a panel with words/terms obtained through authors’ methodology. Also, based on this lists more effective materials for promoting could be elaborated in order to attract students. Future directions in authors' work are to identify the dynamic and trends of the IT Job labor market and its profile for a longer period of time.

Acknowledgement

The analyzed IT Jobs were obtained from the processing of public data displayed on the website https://www.ejobs.ro/.

REFERENCE LIST

[1] Berry, M.W. (2003). *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer, New York.

[2] Chopra, S., & Golab, L. (2018). *Job Description Mining to Understand Work Integrated Learning*. 11th International Conference on Educational Data Mining (EDM), Raleigh.

[3] Espinoza, L.C., Guerrero, A.R., & Agudo, T.N. (2015). *Specializations for the Peruvian Professional in Statistics: A Text Mining Approach*. SIMBig.

[4] Fellows, I. (2018). *Word cloud R pakage*. https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf.

[5] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques*. Morgan Kaufmann.

[6] HAP - Html Agility Pack (2020). *Html Agility Pack*. https://html-agility-pack.net/.

[7] Kino, Y., Kuroki, H., Machida, T., Furuya, N., & Takano, K. (2017). *Text Analysis for Job Matching Quality Improvement*. International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, Marseille, France, Procedia Computer Science, vol. 112, 1523–1530.

[8] Kobayashi, V.B., Mol, S.T., Berkers, H.A., Kismihok, G., & Hartog, D.N.D. (2018). Text Mining in Organizational Research. *Organizational Research Methods, 21*(3), 733–765.

[9] Maceli, M. (2015). What Technology Skills Do Developers Need? A Text Analysis of Job Listings in Library and Information Science (LIS) from Jobs.code4lib.org. *Information Technology and Libraries, 34* (3), 8–21.

[10] Terek, E., Mitic, S., Cvetkoska, V., Vulkonjanski, J., & Nikolic, M. (2018). The influence of Information Technology on job satisfaction and organizational commitment. *Dynamic Relatioships Management Journal, 7* (2), 39–49.

[11] Text mining and word cloud fundamentals in R (2020). *5 simple steps you should know, Statistical tools for high-throughput data analysis*. http://www.sthda.com/english/wiki/text-mining-and-wordcloud-fundamentals-in-r-5-simple-steps-you-should-know.

[12] Witten, I.H., Frank, E., & Hall, M.A. (2011). *Data mining: Practical machine learning tools and techniques*. 3rd edition, San Francisco: Morgan Kaufmann.

[13] Wowczko, I.A. (2015). Skills and Vacancy Analysis with Data Mining Techniques. *Informatics, 2*, 31–49.

LIST OF TABLES

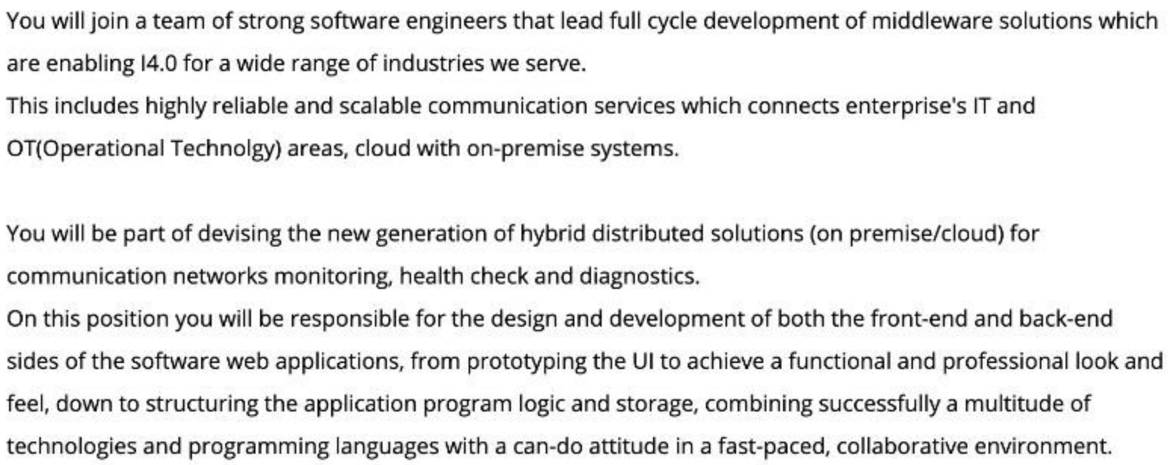
Table 1
IT jobs - April - May 2020 - English- Romanian

Period	IT Jobs English	IT Jobs Romanian
April - May 2020	400 = 77.5%	116 = 22.5%

Table 2
IT jobs - April - May 2020

Month	IT Jobs	It Jobs Percentage
April 2020	174	33.7%
May 2020	342	66.3%

LIST OF FIGURES



You will join a team of strong software engineers that lead full cycle development of middleware solutions which are enabling I4.0 for a wide range of industries we serve.
This includes highly reliable and scalable communication services which connects enterprise's IT and OT(Operational Technolgy) areas, cloud with on-premise systems.

You will be part of devising the new generation of hybrid distributed solutions (on premise/cloud) for communication networks monitoring, health check and diagnostics.
On this position you will be responsible for the design and development of both the front-end and back-end sides of the software web applications, from prototyping the UI to achieve a functional and professional look and feel, down to structuring the application program logic and storage, combining successfully a multitude of technologies and programming languages with a can-do attitude in a fast-paced, collaborative environment.

Figure 1
IT Job Description Example

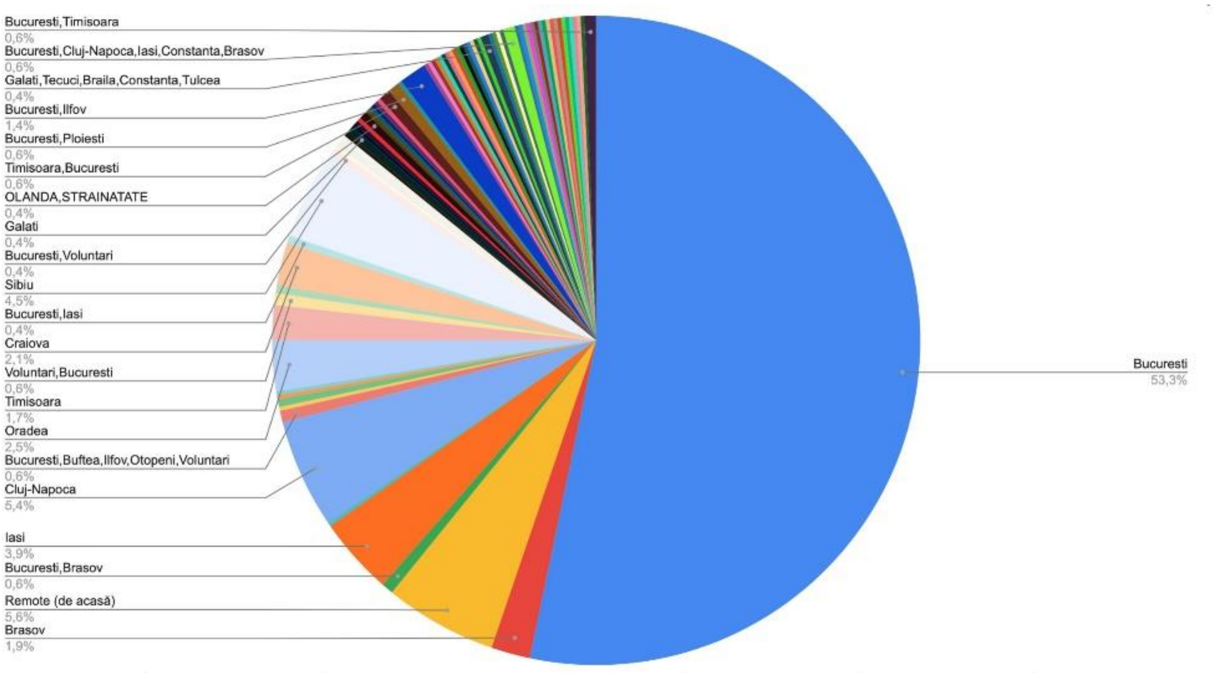


Figure 2
IT Job City Distribution April - May 2020

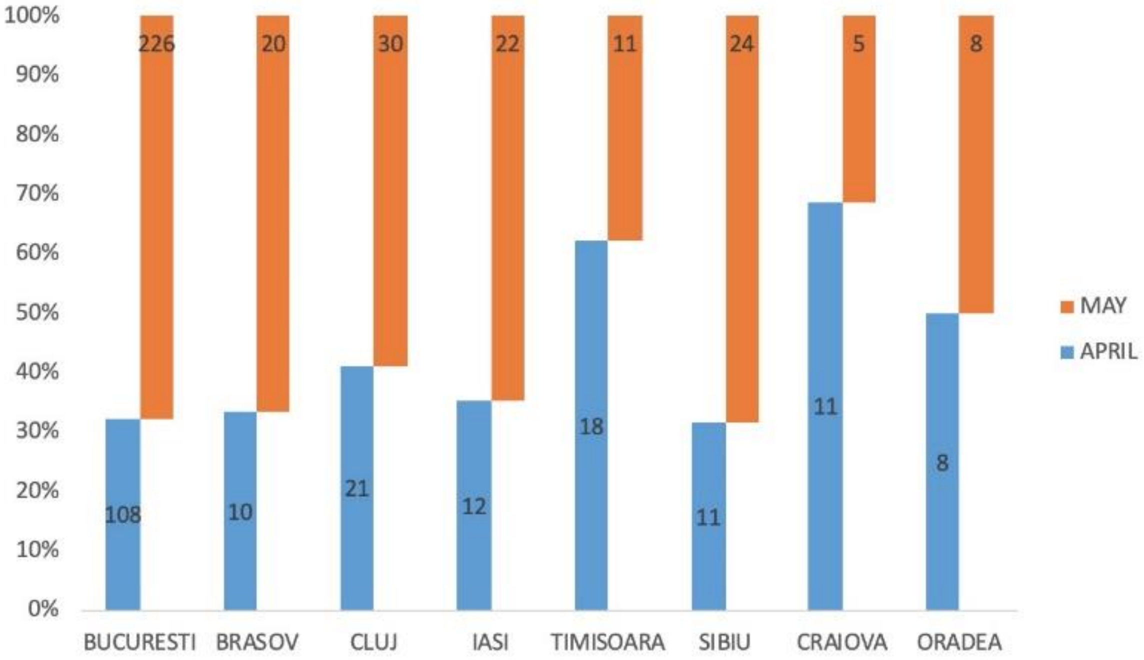
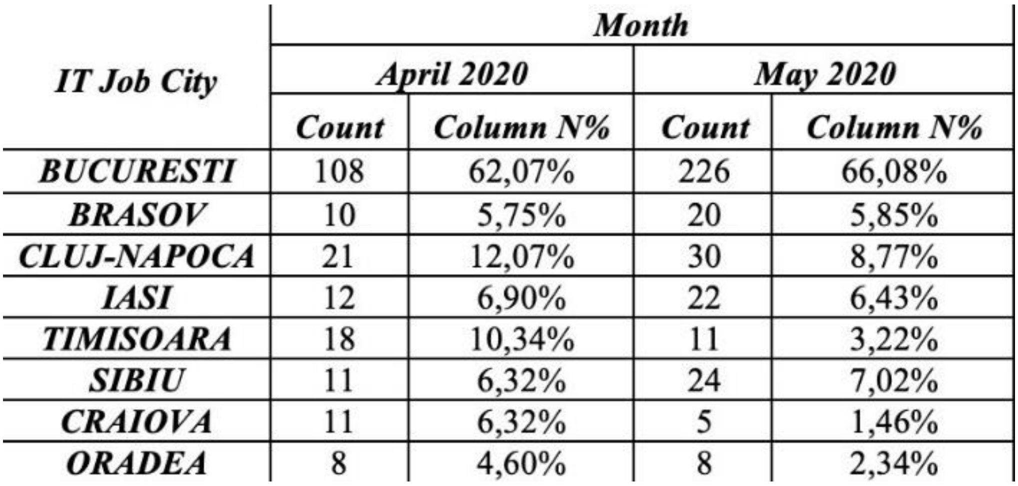


Figure 3
IT Job City Visualization April - May 2020



IT Job City	Month			
	April 2020		May 2020	
	Count	Column N%	Count	Column N%
BUCURESTI	108	62,07%	226	66,08%
BRASOV	10	5,75%	20	5,85%
CLUJ-NAPOCA	21	12,07%	30	8,77%
IASI	12	6,90%	22	6,43%
TIMISOARA	18	10,34%	11	3,22%
SIBIU	11	6,32%	24	7,02%
CRAIOVA	11	6,32%	5	1,46%
ORADEA	8	4,60%	8	2,34%

Figure 4
IT Job City Analysis April - May 2020

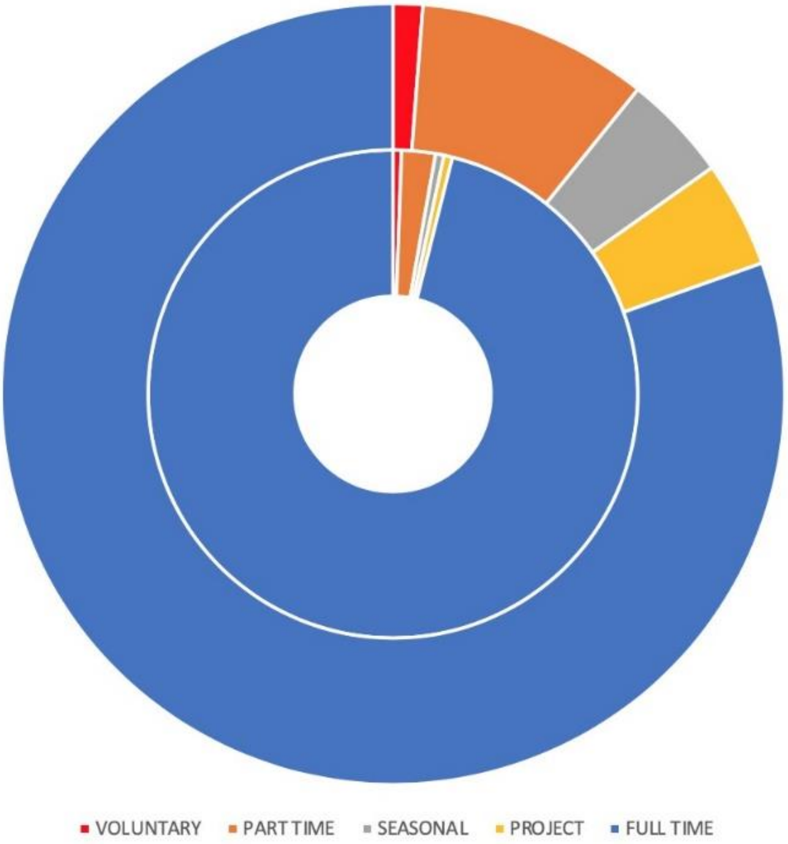
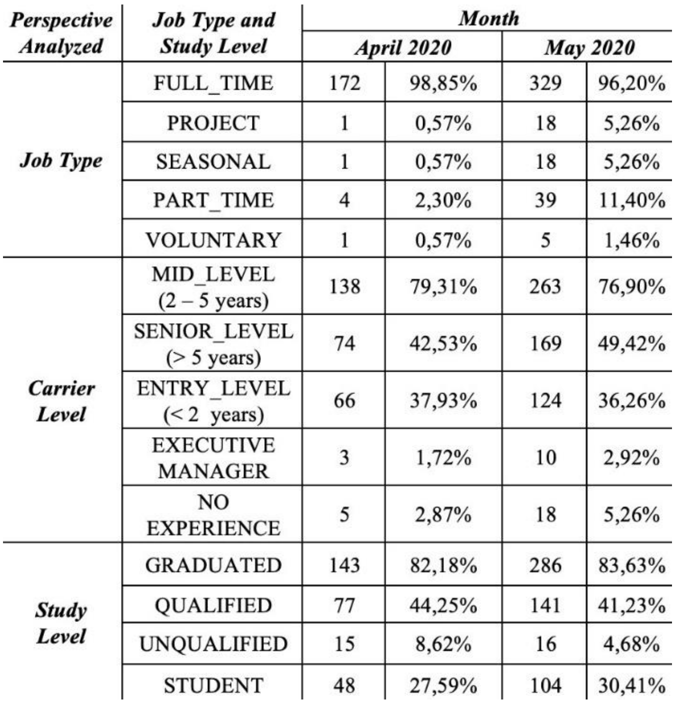


Figure 5
IT Job Type Visualization April - May 2020



Perspective Analyzed	*Job Type and Study Level*	*Month*			
		April 2020		*May 2020*	
Job Type	FULL_TIME	172	98,85%	329	96,20%
	PROJECT	1	0,57%	18	5,26%
	SEASONAL	1	0,57%	18	5,26%
	PART_TIME	4	2,30%	39	11,40%
	VOLUNTARY	1	0,57%	5	1,46%
Carrier Level	MID_LEVEL (2 – 5 years)	138	79,31%	263	76,90%
	SENIOR_LEVEL (> 5 years)	74	42,53%	169	49,42%
	ENTRY_LEVEL (< 2 years)	66	37,93%	124	36,26%
	EXECUTIVE MANAGER	3	1,72%	10	2,92%
	NO EXPERIENCE	5	2,87%	18	5,26%
Study Level	GRADUATED	143	82,18%	286	83,63%
	QUALIFIED	77	44,25%	141	41,23%
	UNQUALIFIED	15	8,62%	16	4,68%
	STUDENT	48	27,59%	104	30,41%

Figure 6

IT Job Type, Career Level, Study Level Analysis April – May 2020

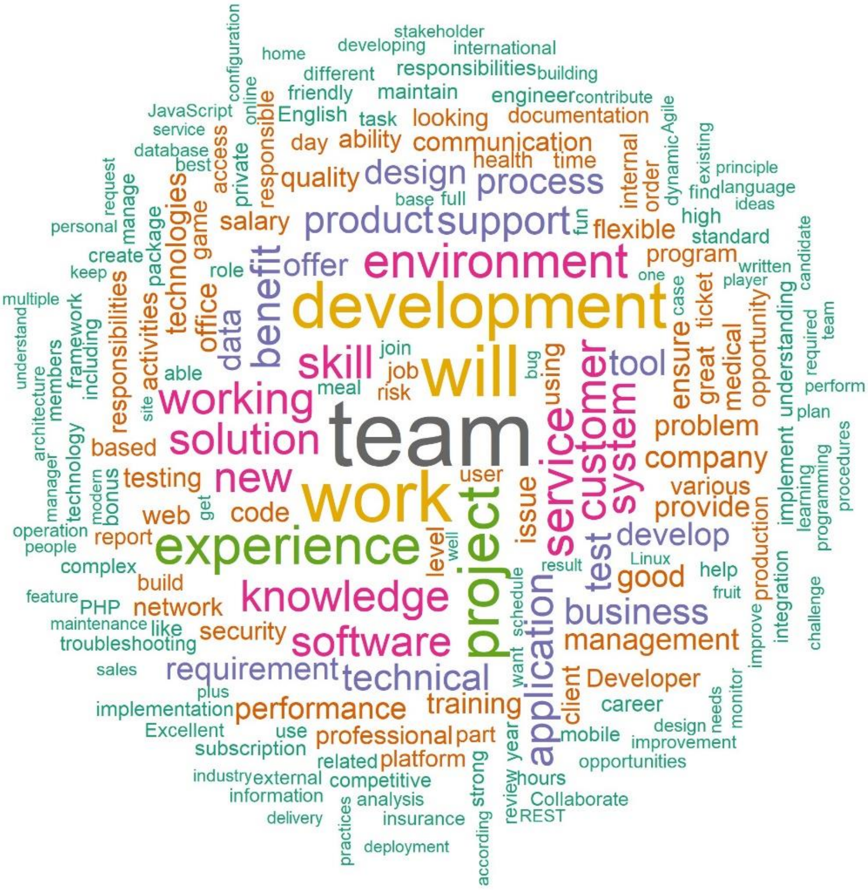


Figure 7

IT Job Description Visualization - English

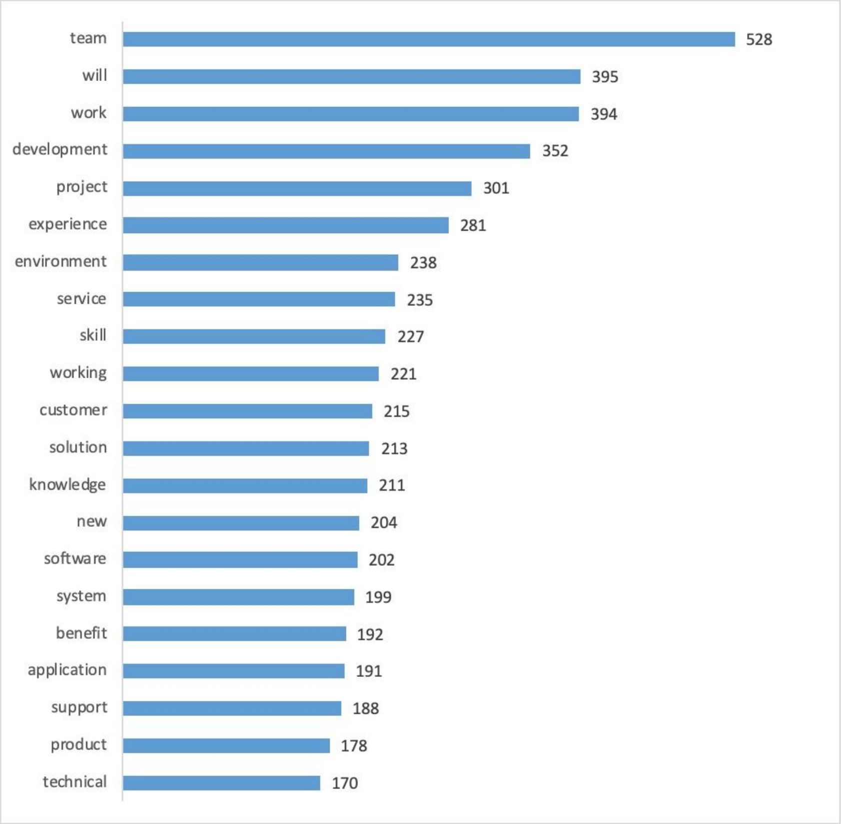


Figure 8
IT Job Description Distribution by words - English

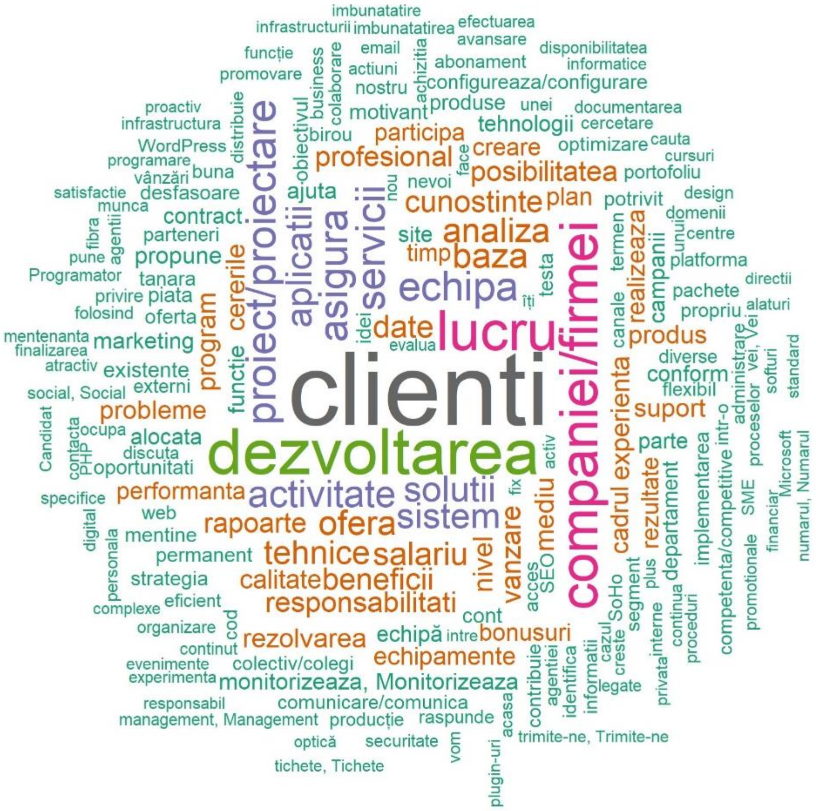


Figure 9
IT Job Description Visualization - Romanian

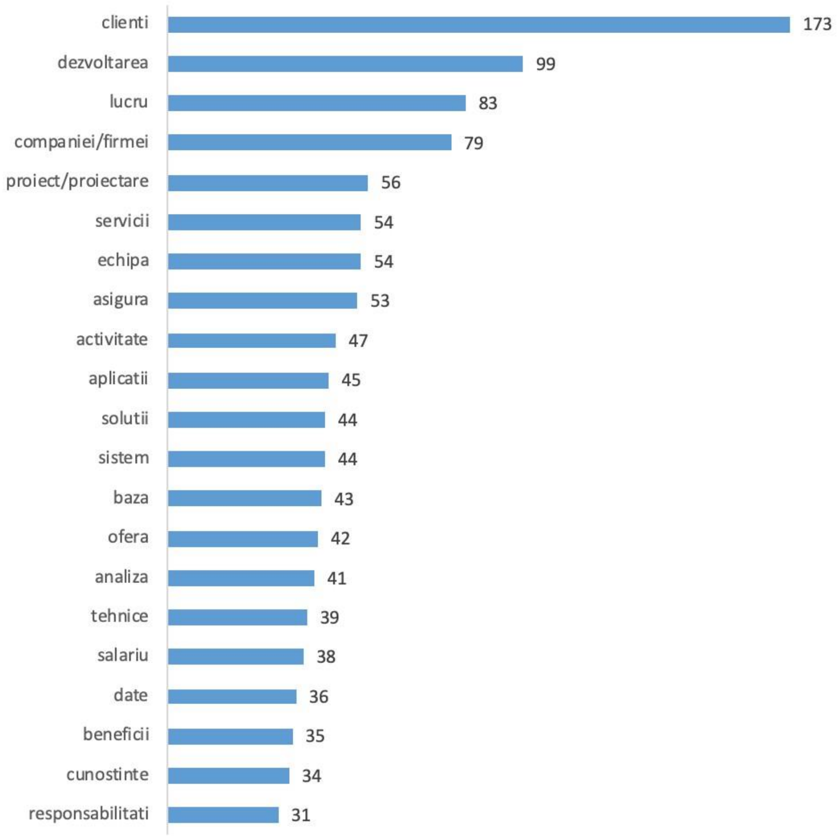
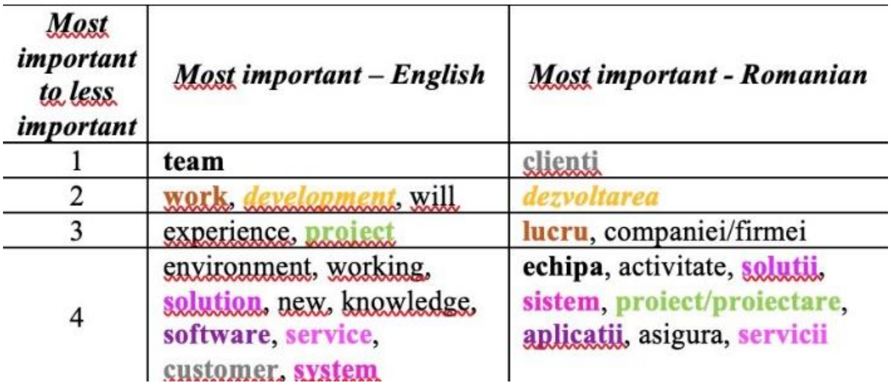


Figure 10
IT Job Description Distribution by words – Romanian



Most important to less important	*Most important – English*	*Most important - Romanian*
1	**team**	clienti
2	**work**, *development*, will	*dezvoltarea*
3	experience, **proiect**	**lucru**, companiei/firmei
4	environment, working, **solution**, new, knowledge, **software**, **service**, **customer**, **system**	**echipa**, activitate, **solutii**, **sistem**, **proiect/proiectare**, **aplicatii**, asigura, **servicii**

Figure 11
IT Job Description Visualization - English, Romanian - Comparison