JOB REQUIREMENTS ANALYSIS WITH DATA MINING TECHNIQUES

Marcu Florentina, *The Bucharest University of Economic Studies, Bucharest, ROMANIA*

ABSTRACT:

Because a large number of job vacancies exist, nowadays, skills research has become the focus of the companies. This study presents a method to identify and analyze the most important requirements for a developer on the Romanian market. The skills needs is a dynamic variable and it depends on many aspects (time, geography, industry, company objectives and so on), having hundreds of thousands of meanings. For monitoring skills needs, as expressed in job advertisements, I used a simple and effective artificial intelligence technology called text mining. To determine knowledge from a given collection of job requirements, I applied selected techniques such as Information Extraction and Information Retrieval whose results have viewed through Word Cloud analysis. All of these could be applied after the data set went through the cleaning process.

KEY WORDS: job requirements, word cloud, text mining, R, frequency matrix

1. INTRODUCTION

In today's society, professional choices are constrained by the labor market demands and the main focus of the companies is to minimize the discrepancies between skills needs and supply. To balance labor markets, the companies invest significant time, financial resources and investigate carefully all changes that occur. Most employers place great emphasis on develop tools that help optimize the assessment and forecast of skills needs. More than that, the selection process of the candidates is undoubtedly one of the top companies priorities. Because there are millions of jobs posted on sites, with different requirements, analyzing these data has become increasingly difficult and, from here, came the need to use data mining tools.

Romanian economy sectors are offering many jobs, the unemployment rate is only 3.90% (December 2019) and the most in-demand are IT jobs. We already know how attractive is the Romanian IT sector, with a diversified and skilled workforce, stimulating environment and competitive salary. However, in this field of activity, the employers' requirements are among the highest and their analysis has become more and more popular on the labor market.

2. TEXT MINING WITH R

Text mining (also known as text analysis) utilizes artificial intelligence (AI) technologies to transform the unstructured text into normalized, meaningful, structured data and actionable information suitable to drive machine learning algorithms. The unstructured texts contain a large amount of information that cannot be used for processing by computers. This problem is rife in analysis of job requirements field. Text mining uses the unstructured data sets to discover secret or unidentified information and focus on extraction of information, clustering, categorization, visualization, summarisation, retrieval information,

analysis of text, data mining and machine learning.

There are five fundamental text mining steps which are shown in the figure below:

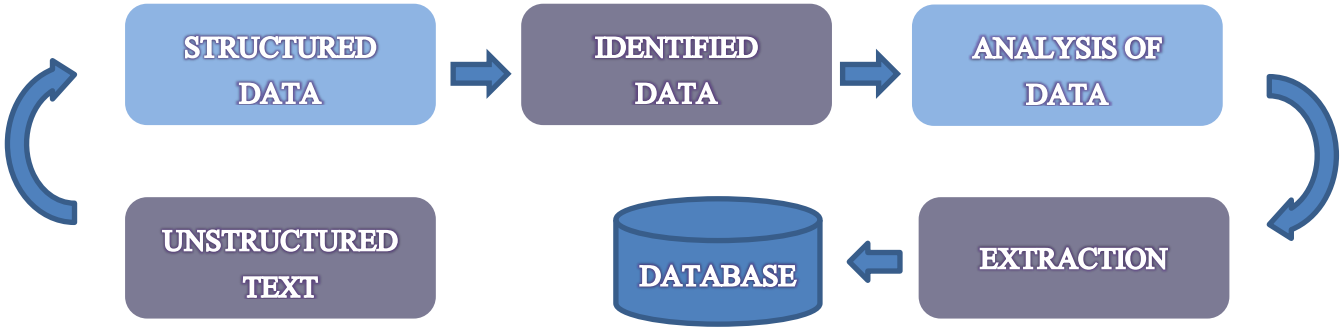

Figure 1. The process of text mining

The data set used in this paper has been retrieved from the eJobs.ro site (unstructured text) and capture the employers' requirements on the Romanian market related to developer jobs. This information has been converted in structured data using the process of data cleaning and then I identify a pattern in the structured data (the frequency of some words). The analysis of data includes: replacing the words that have same meaning, the association measures used to examine the correlation between and the process of removing words that do not add extra information. The penultimate stage, extraction, is performed by the Word Cloud analysis and the storage of the information in the database is an optional step.

In this paper were included 446 developer jobs of different companies, the purpose being to find which are the most important characteristics that an employer expects from all possible applicants for each job.

2.1. THE PROCESS OF DATA CLEANING

I started the model with loading the collected data into the R worksheet. The process of data transformation included the following steps:

1) All words was written in lower case;

2) Removing punctuation marks and white spaces from dataset;

3) Removing numbers and links;

4) Only words between 3 and 20 letters were kept (control parameters);

5) Building the document-term matrix, where on rows we find the documents (number of rows = number of each job requirements document) and on columns all the words that appear in the document (we have as many columns as there are words in the document).

The document is constructed from all identified vacancies (446 vacancies separated by enter). The matrix contains the frequency with which a word appears in each document. For example, for developer requirements matrix, I have 446 rows and 3.352 columns. All of these can be shown in the figure below:

dtm	list [446 x 3352] (S3: DocumentTer	List of length 6
i	integer [16210]	1 1 1 1 1 1 ...
j	integer [16210]	1 2 3 4 5 6 ...
v	double [16210]	1 1 1 1 1 1 ...
nrow	integer [1]	446
ncol	integer [1]	3352
dimnames	list [2]	List of length 2

Figure 2. Document-term matrix

6) The frequency matrix was calculated for the first 30 records (Figure 3. Most common word in job requirements - developer) and can be noted that experience, skills, knowledge,team, ability, development and english are some of the most common words found in developer vacancy, exceeding 150 appearances.

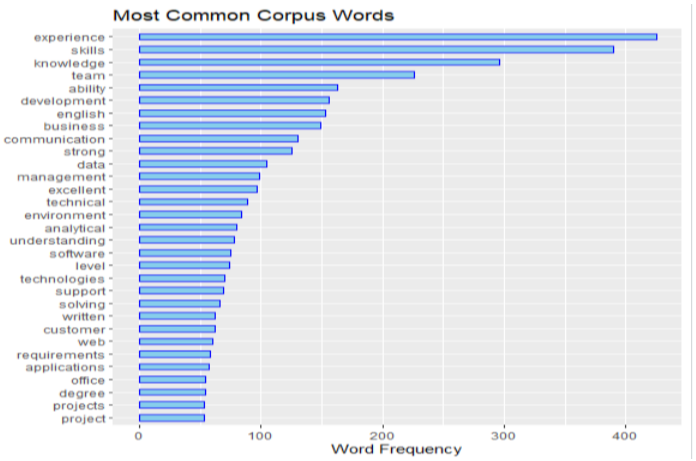

Figure 3. Most common word in job requirements - developer

The initial word cloud consists of the first 100 most common words and is viewed at this step to help us complete the following steps.

Figure 4. The initial Word Cloud

The Figure 4 show that there are words with the same meaning – the word "years" refers to the requirements related to the experience of the candidates. There are also words that do not influence the analysis and do not provide relevant information. These should not appear on the graph and will be eliminated in the next steps of data cleaning process.

7) Replacing the words that have same meaning:
   - office/ms/excel with Microsoft
   - development with develop
   - technologies with technology
   - projects with project
8) Using the association measures to examine the correlation between the words:

```
> findAssocs(dtm, "data",0.5)
$`data`
          big         spark         kafka   participate       adheres centralization
         0.69          0.69          0.67          0.62          0.56          0.56
cloudcontainer devopssupport       enables   exploration          hdfs          hive
         0.56          0.56          0.56          0.56          0.56          0.56
       impala          lake      pipeline   questioning           raw      reusable
         0.56          0.56          0.56          0.56          0.56          0.56
    stability        strict    suggesting   warehousing         asset       airflow
         0.56          0.56          0.56          0.56          0.52          0.51
    analytics  architecting commercializing    downstream         druid        hadoop
         0.51          0.51          0.51          0.51          0.51          0.51
       major       migrate    opensource   partitioning     streaming     threading
         0.51          0.51          0.51          0.51          0.51          0.51

> findAssocs(dtm, "knowledge",0.2)
$`knowledge`
      experience          network            tools        standards
            0.55             0.50             0.48             0.46
      technology             good             unix          english
            0.44             0.44             0.44             0.42
          vision      programming    communication             itil
            0.42             0.41             0.40             0.40
       protocols    methodologies             plus  troubleshooting
            0.40             0.39             0.38             0.38
           git       javascript        technical         computer
            0.37             0.37             0.37             0.37
        domain         approach          areyou  behind\u0094
            0.37             0.36             0.36             0.36
     b\xf6rje          capture          century              ceo
            0.36             0.36             0.36             0.36
  connectivity          ekholm         ericsson            etom
            0.36             0.36             0.36             0.36
           ibm     increasingly     intelligent              iot
            0.36             0.36             0.36             0.36
    javagroovy          leaving          netcool              oss
            0.36             0.36             0.36             0.36
     president programmingscripting          putting           remain
            0.36             0.36             0.36             0.36
```

Figure 5. Terms associations within Job Requirements

The above information provides association between the selected words. The higher the association, the stronger the relationship between the words in the dataset. This method allows us to understand the requirements of companies for IT developers.

For example, we noticed that, within the IT developers posts, the knowledge term relate to network requirements, english, IT protocols, IT methodologies, technical and programming skills, good communication skills, ITIL certification and technologies, such as: HTML, JavaScript, IoT, SQL, Unix and so on. The data skills are requirements related to big data, tools like Apache Kafka, Spark, working in cloud environment, Data warehouse and problems related to the architecture of IT systems.

9) Removing words that do not add extra information to the analysis, such as looking, strong, written, work, advantage, advanced, player, part, people, able, join, using, role, professional, understanding, years, language, similar, minimum, new, least and so on.

All the above transformations were performed by applying specific functions in the analysis environment R. Thus, a clean data corpus was obtained. This data set will be used in the data visualization

subsection to facilitate the interpretation of the most important IT requirements.

2.2. DATA VISUALIZATION

One of the principal objectives of this study was to introduce a method for obtaining information about skills requirements from published jobs. In this field, I illustrate how the content might be visualized to extract detailed information related to the skills in demand. For a good representation of the information, I applied R's *wordcloud* package.

Figure 6. Job Requirements Visualization—Developer jobs

From the figure 6 it is observed that the size of the words corresponds to the frequency of occurrence in the job requirements analyzed. Not surprisingly, the information from the above figure illustrates the importance of experience. Thus, the next most commonly used word is knowledge which means that the potential candidates for the developer position must have extensive programming, computer and technical knowledge. Developer jobs require a large number of technical competences (i.e., PHP programming, oracle technology – SQL, database design and solutions and so on). In addition, a good developer must have excellent communication and analysis skills, work experience, knowledge of English. Other important skills are team working, time management and business knowledge.

3. CONCLUSION

In this study I presented the subject of data mining applied in order to analyze the necessary skills and experience to become developer on the Romanian labor market. For these, I used the text mining process which can be quickly applied in an R environment and I proposed an approach to discover the most important skills for a developer. Through this paper I have offered a detailed overview of competency requirements and may be used as a starting point for a technical examination of the candidate variety by the HR department or in education and training areas. For a faculty, with a technical profile, this application can help teachers to update their subject to the market

demands. In this way, the students will be prepared with current knowledge after graduating.

From this perspective, can be suggested other approaches. Instead of job requirements analysis, can be evaluated the job description, job title or the benefits offered by the companies. Instead of developer jobs analysis, the application can be expanded by analyzing the entire IT domain or IT occupational areas (developer, analyst, administrator, security, engineer and so on) or technologies (SQL, R, Python, Java, ERP, Oracle, SAP and so on). These can also be evaluated according to geographical locations, employer categories, salary ranges and contract types.

REFERENCES

[1] Berry M. W., Kogan J., Text Mining: Applications and Theory, Wiley Publishing House, USA, 2010

[2] Hadley W., R Packages. O'Reilly Media Publishing House, USA, 2017

[3] Hadley W., Garrett G., R for Data Science, O'Reilly Media Publishing House, USA, 2017

[4] Silge J., Robinson D., Text Mining with R. O'Reilly Media Publishing House, U.S., 2017

[5] Tariq S., Text Mining and its Applications. International Journal of Allied Practice, India, 2017 https://www.researchgate.net/publication/323254549_Text_Mining_and_its_Applications

[6] Wowczko I., Skills and Vacancy Analysis with Data Mining Techniques. Institute of Technology Blanchardstown, Dublin, 2015 https://www.mdpi.com/journal/informatics

[7]*** www.eJobs.ro

[8]*** www.datascienceplus.com