Logos Universality Mentality Education Novelty: Economics and Administrative Sciences
ISSN: 2284-5984 | e-ISSN: 2501-0425
Covered in: ERIH PLUS, CEEOL, Index Copernicus, CrossRef, CrossCheck, J-Gate, Google Scholar, Ideas RePeC, Econpapers, Socionet, KVK, WorldCat.

2019, Volume 4, Issue 1, pages: 21-32| doi: 10.18662/lumeneas/10

Employers' Requirements for Data Scientists – an Analysis of Job Posts

Monica Mihaela MAER MATEI[1], **Anamaria Beatrice ALDEA**[2]

[1] PhD, National Scientific Research Institute for Labour and Social Protection, Bucharest, Romania, matei.monicamihaela@gmail.com

[2] Researcher, National Scientific Research Institute for Labour and Social Protection, Bucharest, Romania, anamaria.aldea@incsmps.ro

Abstract: Technological development and innovation are the main drivers of jobs transformations leading to skill mismatch. One very dynamic domain, dealing with these issues is data science. Generally, a data scientist has to work with big data in a scientific and creative manner. To reduce the drawbacks of a sparse matching between educational offer and the new requirements of the labour market is essential to understand real time job market requirements. The most relevant data source for such an investigation is represented by online job market portals. Nowadays, with the increasing digitalisation of society, these portals are considered to improve transparency and signalling in labour markets. Moreover, the potential of the textual vacancy data from Romanian online recruiting platforms has not been exploited up to now. Following these arguments, in order to understand employers' requirements for data science jobs in Romania, we develop an analysis of textual data extracted from job advertisements dedicated to data scientists. Mainly the data analysis will involve the investigation of term frequencies and associations combined with relevant visualization tools. The research will reveal the employers' needs and will support training providers like universities to adapt curricula and training programmes so that they provide what the labour market requires. Moreover, the findings of this research could support young people in making better training choices, signal important trends related to occupations and skills.

Keywords: *labor market; data scientist; text mining; job posts.*

How to cite: Maer Matei, M.M., & Aldea, A.B. (2019). Employers' Requirements for Data Scientists - an Analysis of Job Posts. *Logos Universality Mentality Education Novelty: Economics and Administrative Sciences, 4*(1), 21-32. doi: 10.18662/lumeneas/10

1. Introduction- definitions of data science

The last decades have been characterized by a fast technological development and by innovation growth in many fields. Thus, to keep up, human resources have to hold and gain as many skills as possible in order to increase the organizational productivity, but also to help society evolve in the right direction. In this context of fast changes a new field (or a new science) has emerged: data science.

Data science is a very rewarding field that deals with a fascinating new entity in the data world: big data. However, the term "data science" has appeared before big data. In 1962, the remarkable statistician W. Tukey wrote the book *The Future of Data Analysis* in which he foresaw that a new type of data analysis will rise. The next occurrence of the term was in 1974 in the book of Peter Naur entitled *Concise Survey of Computer Methods*, which contained the first definition of data science: "the science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences" (Voulgaris, 2014).

Although the term appeared more than 50 years ago, the field of data science has become better known at the end of the 1990s, when databases grew larger and the first data science method, called data mining, gained popularity and respect in the scientific community (Voulgaris, 2014). Even if data mining represents an important pillar of data science, this new field will exceed the usage of different analytical tools and methods (Weihs & Ickstadt, 2018), (Cao, 2017b).

From the beginning of the century until now, more and more has been written on this topic and we can say that in the present there are some important definitions and characteristics of this concept that can help us to understand it better. This concept can be seen from several perspectives, so we can say that data science is the modern statistics, a mix of several interdisciplinary fields or a new source of knowledge. Beside these, data science also can provide facilities and new means for the professions related to it and can generate new solutions for business planning (Cao, 2017a).

Data science is a scientific discipline from an interdisciplinary field including elements from statistics, mathematics, informatics, computer science, operations research, as well as the applied sciences (Weihs & Ickstadt, 2018). The formula suggested by Thereat, Cao (2017a) encapsulates this idea:

$$\textit{Data science} = \textit{statistics} + \textit{informatics} + \textit{computing} + \textit{communication} + \textit{sociology} + \textit{management} \mid \textit{data} + \textit{environment} + \textit{thinking},$$

where "|" stands for "conditional on".

From this formula results that statistics and informatics play a central role in data science, but the other fields play an important role too because together lead to the desired and necessary results.

Another definition regarding data science tells that it is an "intelligence science" focused on transforming data into knowledge and wisdom (Cao, 2017b). Moreover, other opinion is that data science is a systematic approach to "thinking with wisdom", "understanding domain", "managing data", "computing with data", "mining on knowledge", "communicating with stakeholders", "delivering products" and "acting on insights" (Cao, 2016).

After all these definitions it is important to say that data science is not merely a set of clever tools, methodologies and know-how. It is a whole new way of thinking about data altogether which requires us to think more systematically and combining an imaginative approach to problems with solid pragmatism (Voulgaris, 2014). Therefore, data science activities are different from the ones involved in standard IT projects, due to its creative and unautomated nature, aiming to discover evidence and indicators for supporting decision-making (Cao, 2016). Thus, the most important steps in data science are: finding structure in data and making predictions (Weihs & Ickstadt, 2018).

Considering the above it is important to admit that we are living in the age of big data and data science which is expected to grow in terms of business value, technology, available knowledge and know-how, and popularity in the years to come (Cao, 2017a, Voulgaris, 2014). In this case, it is very likely that the demand for those who are working in this field, named data scientists, will grow on the labour market.

Data scientists have also been characterized and described over time in many ways in order to better understand this fairly new role in the industry that has grown in popularity since its introduction to the job market. It involves all the different aspects of dealing with data, particularly big data, in an intelligent and very methodical manner, in order to create a useful product. Thus, it can be said that data scientists are the professionals who deal with big data in a scientific, creative and understandable manner. In other words, they are the people that make sense of big data (Voulgaris, 2014).

Other researchers believe that a data scientist is a high-ranking professional with the training and curiosity needed to explore big data as well as a combination of data hacker, analyst, communicator and trusted adviser (Davenport & Patil, 2012). Another opinion is that a data scientist is a professional able to individually cope with most analytical needs of a

company (De Mauro, Greco, Grimaldi & Nobili, 2016). Not only the researchers defined these fairly new profession, but also people having important positions in business related to this field. Thus, Josh Wills, Director of Data Science at Cloudera, said that "a data scientist is a person who is better at statistics than any software engineer and better at software engineering than any statistician". Moreover, Anjul Bhambhri, Vice President of Development for Big Data projects at IBM, considered that data scientists are always inspecting, asking questions, doing "what if" analysis, examining existing theories and procedures.

We saw various opinions about what a data scientist is and what he does, but it is important to know what kind of qualifications and skills he should have to get a job and to perform in this field. These include: Master's or PhD degree in computer science, statistics, mathematics, analytics, data science, informatics, engineering or related fields; background in software engineering; proficiency in statistics, data mining and machine learning techniques; good analytical skills, critical thinking; interest in multidisciplinary studies; ability to implement, maintain and troubleshoot big data infrastructure, such as cloud computing, high-performance computing infrastructure, distributed processing paradigms, stream processing and databases (Cao, 2017a).

2. Skills requirements on the labor market

The main objective of this paper is to investigate labour market requirements for jobs related to data science. We are proposing a solution that aims to extract information form textual data coming from job advertisements using a text mining approach.

Mainly the research undertaken will involve the use of data science and related techniques in order to extract underlying patterns from large collections of data coming from online portals. The increasing digitalisation makes the use of online job market portals an important pillar of recruitment and job search activities. Nowadays online job market portals improve labour market transparency and signalling, supporting a better concordance between labour supply and demand.

Using this information to understand employers' requirements could support training providers like universities to adapt curricula and training programmes so that they provide what the labour market needs.

The importance of this issue resides especially in the current context of technological development and innovations which change the nature of jobs being considered one of the main drivers of skill mismatch (Cedefop,

2019a, b). A sparse matching between skills supply and demand has numerous negative repercussions for individuals and companies. It negatively influences the wages and job satisfaction, and causes hiring difficulties and lower productivity at company level (Cedefop, 2019a, b). Obviously, all these issues have a significant impact on national economy from the following considerations: (i) increase of the unemployment level as a result of inadequate qualification will affect public finances (public costs of unemployment benefits) and will lead as well to broader social consequences such as social exclusion, (ii) poor workforce productivity is correlated to national competitiveness.

Moreover, the relevance of the topic is emphasized by the efforts undertaken in Europe in the last few years with respect to monitoring labour market by using new data, such as:

- The development of the ESCO classification (European Skills, Competences, Qualifications and Occupations) available on an online portal, officially launched in 2017. This is an initiative of European Commission built to describe occupations and skills linked to these occupations. Its contribution to issues such as: matching skills to jobs and training, analysing the labour market, advertising job vacancies is proved by the multitude of projects and platforms relying on it (ESCO annual report, 2019). For example, ESCO framework is used to match people and jobs (Open European Skill Match Maker (Openskimr) (www.openskimr.eu), or to facilitate matching between CVs and vacancies.

- Skills Online Vacancy Analysis Tool for Europe (Skills-OVATE), a tool developed by CEDEFOP that provides dashboards built on data extracted from online job vacancies. This tool is available for 18 countries (Romania is not included), data is extracted from private or public employment job portals, online newspapers *between* 1 July 2018 and 31 March 2019.

- Cedefop's project "*Real-time labour market information on skill requirements: setting up the EU system for online vacancy analysis"* exploits data from online job vacancies to extract real-time knowledge related to labour market and skills in the EU.

- Cedefop's "Digitalisation, AI and the future of work" project is investigating machine learning techniques required to explore the online job vacancy market in order to understand the current skill needs implied by new technologies;

Additionally, there are the scientific papers related to job post investigation using text mining in different domains and for different countries:

- The construction industry was investigated in terms of skills necessary to manage construction projects in US (Gao and Eldin, 2014).

- The required skill set for dealing with big data was compared to the skills needed in business intelligence jobs (Debortoli, Müller and vom Brocke, 2014).

- Transversal competences were investigated for the Polish labour market (Pater, Szkola and Kozak, 2019).

- Labor market dynamics was studied (Hershbein and Kahn, ,2018), (Modestino, Shoag, Ballance, 2016).

- The most demanded occupations in the modern job market were identified (Karakatsanis et al. , 2017)

- IT vacancies were studied in Ireland (Wowczko, 2015)

- Comparing the competence demand of the Helsinki metropolitan area labour market with the supply of competence from universities (Ketamo, Moisio, Passi-Rauste & Alamäki. 2019)

- The skills set needed in the financial sector (Lavrinenko & Shmatko, 2019)

- The skill needs for early career researchers (Maer, Mocanu, Zamfir & Georgescu, 2019)

Moreover, there are few articles evidencing previous work related to the alignment between acquired university curriculum outcomes and required market skills:

- A framework referred to as Align My Curriculum (AMC) (Almaleh, Aslam, Saeedi & Aljohani, 2019) which enables the classification, comparison and visualization of online computing jobs and computing curricula.

- The results of a content analysis of online job advertisements were compared to IS curriculum (Woolridge, 2016).

- An investigation of the German IS curricular module descriptions and offers for IS job starters (Föll & Thiesse, 2017).

- For Romania, there are not evidences of such findings. However, there is a report initiated by CEDEFOP which investigates the possible data sources (Cedefop, 2018a, b) that could be used for understanding job market requirements.

3. Method and Data

The findings presented within this study, are based on the information extracted from posts containing "data science" or "data scientist" in their titles, in the period 29.10.2019 - 13.11.2019, on the

Romanian labour market. Job requirements in terms of skills and knowledge were collected from two online recruitment platforms: Bestjobs.eu and Hipo.ro. There were 15 posts satisfying these criteria. Therefore, a collection of 15 documents representing descriptions of the requirements associated to each vacancy serves as the source of information in our investigation. Generally, in the text analysis literature, a collection of documents is known as a corpus. Text analysis was performed with tm library in R (Feinerer, 2008), (Meyer, Hornik, & Feinerer, 2008) and mainly used for text summarization. The two main characteristics of a corpus are the lexicon and the size. The lexicon or the dictionary is represented by all unique words in a specific corpus meanwhile the corpus size gives the total number of words included in the documents. In order to extract relevant information from these texts, data transformation is mandatory. In this stage, we need to build a structured representation of the corpus. The operations undertaken will produce a matrix known as document- term matrix containing frequencies. The rows of the matrix are the documents included in the corpus and the columns are the terms encountered in the documents (Srivastava & Sahami, 2009). The initial matrix is built after performing some preliminary cleaning operations such as: deleting extra white spaces, eliminating conjunctions and prepositions (stop words), removing punctuation and numbers and converting to lower case. The final dimension of the corpus as presented in Table 1 is obtained after performing also the following operations:

- Discarding common words specific to the job posts, not relevant for our analysis, even if they are encountering high frequencies. For example terms like "skill", "ability", "proficiency", "looking" were dropped in this stage.

- Replacing the plural form of some words with their singular form. For example "teams" was changed into "team" or "visualizations" was substituted by "visualization". In the same logic, certain suffixes were dropped, keeping for example "statistics" instead of "statistical" or "analytic" instead of "analytical")

All these operations are undertaken in R using tm library (Feinerer & Hornik, 2013).

Table 1. Dimension of the corpus

	Vocabulary size	Corpus size
Before Data Cleaning	1199	3203
After Data Cleaning	887	2240

4. Results

The frequency matrix was analyzed to extract the labor market requirements for data science jobs. The findings are synthesized through word clouds, a visual instrument pointing up the terms encountering the highest frequencies in the job advertisements. The frequency of a specific word is given by the sum of the column it represents in the frequency's matrix obtained after conducting the data cleaning operations. The word cloud we have presented in this paper uses the top 30 most frequent words. For a better representation of the words, we have decided to remove the term "data" from the matrix, acting as an outlier given its extremely large frequency.

Our investigation reveals that in Romania, employers looking for data scientists require a mix of skills which is consistent to the formula defining data science (Cao, 2017a) discussed in the first section of the paper. Therefore, in order to become a data scientist, you must have the ability to extract knowledge from data using IT skills, statistics and analytical skills. The word cloud presented in Figure 1, also emphasizes specific aspects related to the required programming skills. Many of the job posts are mentioning that candidates should have experience with tools like Rstudio, Python or Spark.

According to the posts we have included in this analysis, the employers need solutions offered by techniques and algorithms from machine learning.

Figure 1. Employers requirements for data scientists

However, these solutions should be provided by specialists able to understand the business processes and flows. This is why we consider that this jobs are suited for graduates of inter-disciplinary programmes involving economics, computer use, use of software for data programming and statistics.

Another skill, underlined by our output, that data scientist should possess is related to data visualization. In the job advertisements are mentioned the specific tools required to perform this task.

It is important to mention that the term "management" is also associated to data or database so it is not necessarily associated to management skills but rather to IT or data processing.

Apart from these technical skills, communication skills are also needful for performing data science jobs. Related to good communication and collaboration are the teamwork skills, also revealed by our representation in Figure 1. We can include in this last category of abilities valued by employers, the one announced by the occurrence of the term "English". Therefore, data scientists are also expected to be fluent in English.

The main limitations of this study are related to the insufficient time span covered by our data collection process. This issue involved a reduced number of job posts to analyze. We consider that by expanding the investigation on a significantly larger collection of documents we could extract valuable information related to specific tools, methods or algorithms required by the labour market for data science positions. In our future research, in order to increase the data volume, we plan to develop a solution to collect data automatically from posts where the job titles should be about: computer scientist, data analyst, data scientist, systems analysts.

5. Conclusions

This study employs a text analysis approach to find the qualifications and skills required from jobseekers aspiring to work in the data science field. Coupled with the evolution of technology, job market requirements undergo frequent changes. Hence, for a good matching between skills supply and demand, it is necessary that the educational offer keep up with these transformations. Our findings could help training providers like universities to adapt curricula and training programmes so that they provide what labour market needs. The analysis of textual data extracted from job advertisements could also support young people in making better training choices. Moreover, from a scientific point of view, the potential of the textual vacancy data from Romanian online recruiting platforms has not been exploited up to now, which offers the opportunity of innovative research. Our investigation in this field is in the incipient stage but our results are in concordance with the data science literature. These preliminary results show that data mining and processing skills are essential for a data science career. IT skills have also been highlighted by the undertaken analysis. Additionally, the word cloud representation allows the identification of the specific tools that are required such as Rstudio, Python or Spark. As in any other field, apart from these technical skills, good communication and teamwork skills are valued by the employers. This evidence is explicitly revealed by our findings.

Acknowledgement

This paper has been elaborated in the NUCLEU Program 19N/2019, funded by Romanian Research and Innovation Ministry, project PN 19130303.

References

Almaleh, A., Aslam, M. A., Saeedi, K., & Aljohani, N. R. (2019). Align My Curriculum: A Framework to Bridge the Gap between Acquired University Curriculum and Required Market Skills. *Sustainability*, *11*(9), 2607.

Cao, L. (2017a). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, *50*(3), 43.

Cao, L. (2017b). Data science: challenges and directions. *Communications of the ACM*, *60*(8), 59-68.

Cao, L. (2016). Data science: Nature and pitfalls. *IEEE Intelligent Systems*, *31*(5), 66-75.

CEDEFOP (2019a). Online job vacancies and skills analysis, A Cedefop pan-European approach, ISBN: 978-92-896-2850-1 doi:10.2801/097022

CEDEFOP (2019b). The online job vacancy market in the EU: driving forces and emerging trends. Luxembourg: Publications Office. Cedefop research paper; No 72. http://data.europa.eu/doi/10.2801/16675

CEDEFOP (2018a). Real-time labour market information on skill requirements: setting up the EU system for online vacancy analysis.

CEDEFOP. (2018b). Mapping the landscape of online job vacancies. Background country report: Romania,http://www.cedefop.europa.eu/en/events-and-projects/projects/big-data-analysis-onlinevacancies/publications

Davenport, T. H., & Patil, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century-A new breed of professional holds the key to capitalizing on big data opportunities. But these specialists aren't easy to find—And the competition for them is fierce. *Harvard Business Review*, 70.

De Mauro, A., Greco, M., Grimaldi, M., & Nobili, G. (2016). Beyond data scientists: a review of big data skills and job families. *Proceedings of IFKAD*, 1844-1857.

Debortoli, S., Müller, O., & vom Brocke, J. (2014). Comparing business intelligence and big data skills. *Business & Information Systems Engineering*, *6*(5), 289-300.

Feinerer, I., & Hornik, K. (2013). tm: Text Mining Package. R package version 0.5-10. *2014-04-10]. http://CRAN. R-project, org/package= tm*.

Feinerer, I. (2008). An introduction to text mining in R. *The Newsletter of the R Project Volume 8/2, October 2008*, *8*, 19.

Föll, P., & Thiesse, F. (2017). Aligning IS Curriculum with Industry Skill Expectations: A Text Mining Approach.

Gao, L., & Eldin, N. (2014). Employers' expectations: A probabilistic text mining model. *Procedia Engineering*, *85*, 175-182.

Hershbein, B., & Kahn, L. B. (2018). Do recessions accelerate routine-biased technological change? Evidence from vacancy postings. *American Economic Review*, *108*(7), 1737-72.

Ketamo, H., Moisio, A., Passi-Rauste, A., & Alamäki, A. (2019). Mapping the Future Curriculum: Adopting Artificial Intelligence and Analytics in Forecasting Competence Needs. In *Proceedings of the 10th European Conference on Intangibles and Intellectual Capital ECIIC 2019*. Academic Conference Publishing International.

Karakatsanis, I., AlKhader, W., MacCrory, F., Alibasic, A., Omar, M. A., Aung, Z., & Woon, W. L. (2017). Data mining approach to monitoring the requirements of the job market: A case study. *Information Systems*, *65*, 1-6.

Lavrinenko, A., & Shmatko, N. (2019). Twenty-First Century Skills in Finance: Prospects for a Profound Job Transformation. *Форсайт*, *13*(2 (eng)).

Maer-Matei, M. M., Mocanu, C., Zamfir, A. M., & Georgescu, T. M. (2019). Skill Needs for Early Career Researchers—A Text Mining Approach. *Sustainability*, *11*(10), 2789.

Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in R. *Journal of statistical software*, *25*(5), 1-54.

Modestino, A. S., Shoag, D., & Ballance, J. (2016). Downskilling: changes in employer skill requirements over the business cycle. *Labour Economics*, *41*, 333-347.

Pater, R., Szkola, J., & Kozak, M. (2019). A method for measuring detailed demand for workers' competences. *Economics: The Open-Access, Open-Assessment E-Journal*, *13*(2019-27), 1-30.

Srivastava, A. N., & Sahami, M. (2009). *Text mining: Classification, clustering, and applications*. Chapman and Hall/CRC.

Voulgaris, Z. (2014). *Data scientist: the definitive guide to becoming a data scientist*. Technics Publications.

Wowczko, I. (2015, December). Skills and vacancy analysis with data mining techniques. In *Informatics* (Vol. 2, No. 4, pp. 31-49). Multidisciplinary Digital Publishing Institute.

Woolridge, R. W., & Parks, R. (2016). What's In and What's Out: Defining an Industry-Aligned IS Curriculum Using Job Advertisements. *Journal of Higher Education Theory & Practice*, *16*(2).

Weihs, C., & Ickstadt, K. (2018). Data science: the impact of statistics. *International Journal of Data Science and Analytics*, *6*(3), 189-194.