eBISS 2019
Berlin , 1 – 5 July 2019
Université Libre de Bruxelles

EMPLOYABILITY SKILLS AT THE AGE OF BIG DATA: EXPLORING DATA SCIENTISTS' REQUIREMENTS IN THE DIGITAL LABUR MARKET

FRANCESCO SMALDONE
PhD Student in Big Data Management
Department of Business Sciences, Management & Innovation Systems (DISA-MIS) - University of Salerno
Via Giovanni Paolo II, 132, 84084, Fisciano (SA), Italy
fsmaldone@unisa.it

Scan me

UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI SCIENZE AZIENDALI MANAGEMENT & INNOVATION SYSTEMS

INTRODUCTION

The continuous and growing change and globalization are plunging individuals into a system that requires continuous analysis and selection of increasingly massive data sets. These raw data, produced by individuals, are increasingly extended in terms of volume and digital nature, are forcing many companies, in addition to public administrations, to hire professionals who can read and interpret them: data scientists. This overabundance of data is having a strong impact on work world and decision-making dynamics and continues to extend well beyond the boundaries of the business, touching on other important realities such as education, public health, sport, science and management (Provost & Fawcett, 2013). Thus, this happens because the type of analysis required of these specific professional figures involves a careful reading of digital bits of intelligence, starting with Tweets and Facebook Posts, continuing with web scraping in order to analyze and index data and ending with the field, political, financial and labor market analyses. Moreover, the Analytics market represents the main field of application of these new professional figures. As reported from the Istituto Nazionale di Statistica (https://www.istat.it), only in 2016, in Italy, this market grew by 15%, reaching a total value of 905 million euros. Business Intelligence remains the predominant component with 722 million (+9% in one year), but the Big Data part (183 million) has grown by as much as 44%. Much of the market is composed of large companies (87% of total expenditure), with the remaining 13% coming from SMEs. The scientific and working context has made it necessary to operate through Big Data primarily due to the clear transition from Analytics 1.0 to 2.0 of Big Data Analytics, requiring more and more hard and soft skills in order to face new contests and technologies.

BACKGROUND OVERVIEW

Nowadays, the increasing attention to the continuous amount of unstructured, semi-structured and structured data, obtained from websites and social networks requires at the age of Big Data professionals who are flexible, dynamic and versatile within the modern business scenarios (Gehl, 2015). In the literature, there are many case studies of the use of Data Scientists (DSs) in the various market sectors related both to industrial production and the provision of services. Usually, DSs are employed mainly in mining data and analyze stored target data to provide diagnostic, prescriptive, perspective and predictive statistics in order to support the corporate management in creating value from business analytics (Vidgen et al., 2017). DSs' actual role in a corporation is to conduct independent research and analyses on enormous volumes of unstructured, semi-structured and structured data from various internal and external sources (Kim et al., 2016). Structuring and interpreting the extracted data, DSs implement advanced analytics programs, statistical techniques, and methodologies, including machine learning, deep learning, data mining, and text mining, to manipulate data and use these data in modeling (Van der Aalst, 2016; Witten et al., 2016; Ratner, 2017). Analyzing data, DSs thoroughly cleaning and uniting the knowledge extracted from the information to discard anything that is not useful to the task. DSs are mainly employed in mining trends, opportunities, and hidden weaknesses within the data (Wang et al., 2016). As suggested from Kim et al. (2016), to DSs are required, in addition to a high quality education, analytic problem-solving skills, and other hard skills are required such as fundamental competencies and an effective communication, other than ability to work-in-team, intellectual and technological attitudes and a deep knowledge of the industry in which they are employed, further to other earned soft skills. Other studies in the literature highlighted the range of skills averagely owned by DSs, an example comes from Costa & Santos (2017), exploring the competence framework and the skills framework over DSs and underlining the strong component derived from soft skills, absolute value added for these professional figures usually raised up in a hard skills' educational framework or multi-disciplinary oriented courses combining subjects from computer science, statistics and business intelligence. As reported by the United States Bureau of Labor Statistics (https://www.bls.gov), the demand for computer and information research scientists is expected to rise 19 percent by the year 2026. These figures highlight that DSs' employability is rising faster than the average for all the other professions, recording about 5.400 new job open positions projected over the decade, including data-mining services. According to PayScale (https://www.payscale.com/research/US/Job=Data_Scientist/Salary), the average salary for data scientists in the United States is about $91,000 per year; salaries vary on the base of many factors, as for example location, firm, industry, experience, and skills. These considerations have conducted the following research questions:

RQ_1: What are the most requested skills for Data Scientists of entrepreneurs offering jobs?
RQ_2: What are the key disciplines to which these skills refer?
RQ_3: Is there a relationship of dependency between the skills required for data scientists?
RQ_4: How can these skills be configured in a social network?

METHODOLOGY

The first phase of the methodology involves the analysis of the text using Text Mining (TM) in order to analyze, clean, organize and extract knowledge from the text, to then proceed to an analysis of the occurrences and produce a graphic representation by means of a word cloud. Text Mining represents the most complex extension of Data Mining, it has developed since the mid-1990s and it is essential above all to companies and institutions to deal with the excess of information, due, in turn, to the growing availability of IT resources (such as electronic dictionaries). Text Mining simultaneously connects Information Retrieval and Information Extraction operations and is structured in different phases. (1) The pre-processing phase of texts (in which computer science prevails) consisting of finding sources of texts from the web or Intranet, in their formatting (e.g. transformation into XML) and in the constitution of the document warehouse. (2) Stage of lexical processing (in which Linguistics prevails) consisting in recognizing words (with use of dictionaries and knowledge bases, semantic networks, sensigraphs or other), identifying already known keywords or concepts (with use of rules and ontologies), perform lemmatization (recognition of the main parts of speech, especially nouns, adjectives and verbs). This is not a necessary phase for all applications, because sometimes the linguistic processing of the text is not performed. (3) Real Text Mining phase (in which Statistics and Data Mining techniques play a crucial role) consisting of one or more of the following steps:

- Automatic categorization of documents for subsequent retrieval of information;
- Search for entities (terms) in texts that are also multilingual, therefore also independently of the language of origin of the terms, this presupposes the availability and alignment of specific linguistic resources in the different languages investigated;
- Queries in natural language, interpreted by NLP processes also based on artificial intelligence algorithms.

The second phase involves the implementation of a Cluster Analysis (CA), first conducted with a complete method and obtaining a cluster dendrogram as output, and then using a density-based approach (DBSCAN), proposed for the first time by Ester et al. (1996), obtaining a spatial bidimensional map for the visualization.
The third step provides for the visualization of the associations between the skills through an association plot, to show the results both by numbers and by a heat map (Wei et al., 2017).

The fourth and last phase of the methodology foresees the visualization of the skills through a network (Weinberg, 1962; Granovetter, 1983) that highlights the main interconnections between these, for the network are supplied average of the bonds, standard deviation, direct and indirect density indices, and the centrality measures. Finally, data collection has been conducted on the American job ads listing on the Indeed portal. The choice of the American market (US) is motivated by the fact that, in relation to the keyword "data scientist" for job vacancy searches on the Indeed research portal, the American market offers around 6000 more job listings than the English, French and Italian markets (9000 observations in the American market against 3000 observations in the English market, 2500 observations in the French market and just over 800 for the Italian market)

RESULTS

The sample was composed of 395 job ads extracted from the Indeed portal in the American (US) market recording 1.9kk characters in 383 documents, excluding invalid cases. After conducting text analysis trough R software (https://cran.r-project.org), where the text has been cleaned, organized and analyzed in order to extract knowledge, terminological combinations have been obtained in form of n-grams trough weka function and a tokenized Document-Term Matrix (DTM) with a 383x31198 dimension has been built, recording 15086/6705474 non-sparse/entries. In order to extract skills from the text it was necessary to reduce the original sparsity of the matrix (98%); sparsity was accurately reduced until 33%. The maximum term length was equal to 65 and, the matrix was weighted on the term frequencies. Thus, skills have been extracted and the main frequencies have been reported in the bar plot in Fig. 1. Furthermore, analyzed the frequencies of the corpus a weighted and tokenized word cloud were built in order to highlight the main skills and prerequisites required to a data scientist by American entrepreneurs, the word cloud is reported in Fig. 2. Text analysis enabled to reply to the first research question (RQ1): responding, the main skills required to a data scientist in the American (US) market are the following: big data management, algorithms, artificial intelligence, data mining, data sets, software developing, business solutions, machine learning, analytics, big data analytics, r, python, soft skills, hard skills, communication skills, data analysis, work-in-team, data sources, years experience, data processing, data science, bachelor degree, business solutions, deep learning, computer science, data technologies, business intelligence, information technology, data scientist, data management, software engineering. Aimed to respond to the second research question (RQ2), cluster analysis was conducted, both with Complete Method and Density-based method (DBSCAN), obtaining three clusters provided in a dendrogram visualization (Fig. 3) and in a spatial bidimensional map (Fig. 4). Responding to the research question we can almost say that there are three disciplines subordinated to required data scientists' skills, respectively: big data management, computer science, and data science. Once features and labels were extracted and organized, associations have been tested in order to understand the dependency relation trough statistical characters (RQ3). Mixed methods have been employed for the association analysis and represented through an association plot. Associations values and heatmap for visualization are provided in Fig. 5, labels and figures in the plot are following hierarchical clustering disposition. Finally, social network analysis (SNA) representation has been employed in order to respond to the last research question (RQ4). The skills network has been represented in two ways: with undirected tie method (Fig. 6), with directed tie method (Fig. 7), with the implementation of Watts-Strogatz algorithm (Watts & Strogatz, 1998). The network presented a directed density equal to .518, showing a medium density in the graph, an undirected density equal to .753 showing a medium-strong density utilizing this method, a ties' mean of 16.16 and a standard deviation of 2.71 showing a good variability. The most central skills using normalized degree centrality were data technologies (90%), big data analytics (86.6%), deep learning (83%), machine learning (83%) and data mining (81%). The most between skill was data management (st.bet = 87,4%) and the closest skill was computer science (st.clos = 83.2%).

CONCLUSION, IMPLICATIONS AND FUTURE DIRECTIONS

This kind of skills mapping does not exist in the literature, as in the definition of skills useful to employability has never analyzed a specific professional figure, preferring the classification of skills in communicative, employment, hard and soft. Analyzing the skills mainly required by data scientists can have different applications, both in terms of business sciences and management and in terms of analysis techniques that can be implemented in order to carry out an increasingly effective mapping. Mapping the main skills required by a professional figure allows companies to support the decision when selecting and recruiting human resources. Together with statistics, data analysis and software engineering it is possible to realize a real decision support system for the HR management, which allows to streamline the decisional and organizational processes having at its disposal a real dossier of the candidates which can be classified more easily by using digital tools up to the semi-automatic selection of resources by defining thresholds in the analyzed values. The prototype of the DSS is provided online as an interactive heatmap, available to you scanning the QR code at the top of the page on the right. In the future, we will try to test a semi-automatic analysis of resources with a decision based on the threshold values obtained in relation to the parameters on a larger and more significant sample of job advertisements.

MAIN REFERENCES

Addo-Tenkorang, R., & Helo, P. T. (2016). Big data applications in operations/supply-chain management: A literature review. Computers & Industrial Engineering, 101, 528-543.
Akter, S., Wamba, S. F., Gunasekaran, A., Dubey, R., & Childe, S. J. (2016). How to improve firm performance using big data analytics capability and business strategy alignment?. International Journal of Production Economics, 182, 113-131.
Chen, W., Quan-Haase, A., & Park, Y. J. (2018). Privacy and Data Management: The User and Producer Perspectives. American Behavioral Scientist, 0002764218791287.
Costa, C., & Santos, M. Y. (2017). The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age. International Journal of Information Management, 37(6), 726-734.
Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).
Gehl, R. W. (2015). Sharing, knowledge management and big data: A partial genealogy of the data scientist. European journal of cultural studies, 18(4-5), 413-428.
Granovetter, M. (1983). The strength of weak ties: A network theory revisited.
Kim, M., Zimmermann, T., DeLine, R., & Begel, A. (2016, June). The emerging role of data scientists on software development teams. In Proceedings of the 38th International Conference on Software Engineering (pp. 96-107). ACM.
Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big data, 1(1), 51-59.
Ratner, B. (2017). Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data. Chapman and Hall/CRC.
Read, K. (2016). Book Review: The Accidental Data Scientist. Journal of the Canadian Health Libraries Association/Journal de l'Association des bibliothèques de la santé du Canada, 37(1).
Schoenherr, T., & Speier-Pero, C. (2015). Data science, predictive analytics, and big data in supply chain management: Current state and future potential. Journal of Business Logistics, 36(1), 120-132.
Storey, V. C., & Song, I. Y. (2017). Big data technologies and management: What conceptual modeling can do. Data & Knowledge Engineering, 108, 50-67.
Tudoran, R., Costan, A., & Antoniu, G. (2016). Overflow: multi-site aware big data management for scientific workflows on clouds. IEEE Transactions on Cloud Computing, 4(1), 76-89.
Van der Aalst, W. (2016). Data science in action. In Process Mining (pp. 3-23). Springer, Berlin, Heidelberg.
Vidgen, R., Shaw, S., & Grant, D. B. (2017). Management challenges in creating value from business analytics. European Journal of Operational Research, 261(2), 626-639.
Wang, H., Xu, Z., Fujita, H., & Liu, S. (2016). Towards felicitous decision making: An overview on challenges and trends of Big Data. Information Sciences, 367, 747-765.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world 'networks. nature, 393(6684), 440.
Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., & Zemla, J. (2017). Package 'corrplot'. Statistician, 56, 316-324.
Weinberg, L. (1962). Network analysis and synthesis (Vol. 249). New York: McGraw-Hill.
Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Figure 1: Bar plot with skills frequencies count.

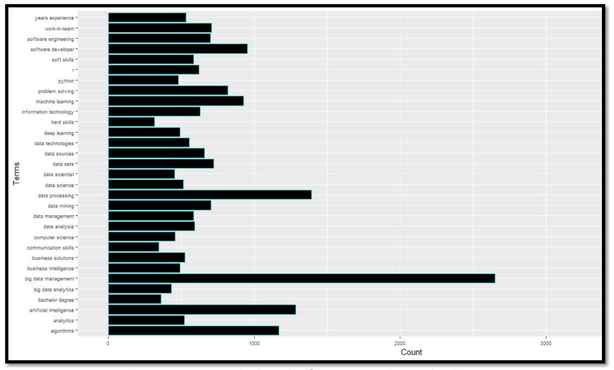

Figure 2: Word cloud after sparsity reduction.

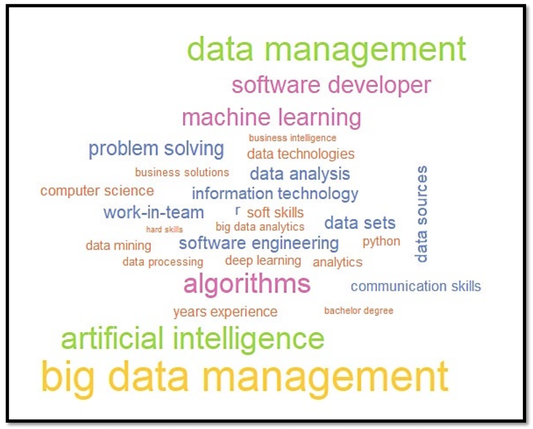

Figure 3: Cluster dendrogram with complete method.

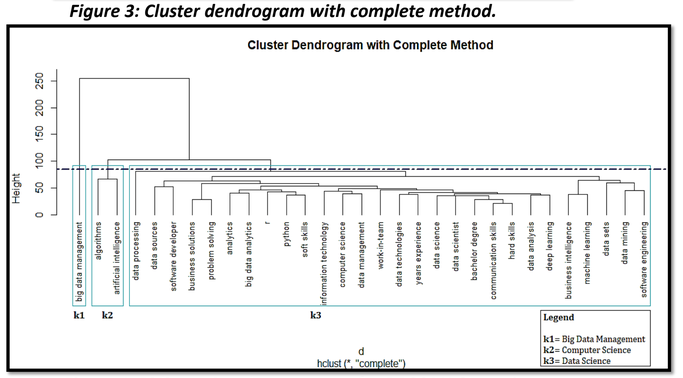

Figure 4: Bidimensional map output for DBSCAN Clustering.

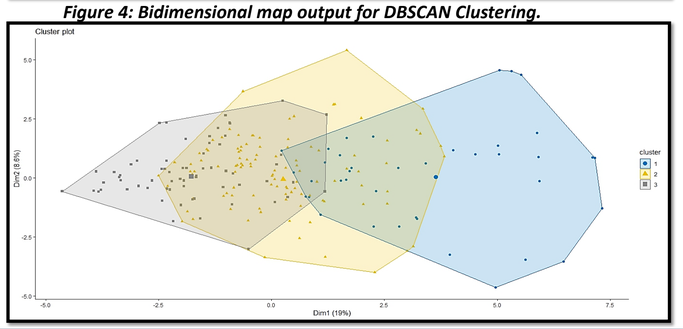

Figure 5: Mixed-methods association plot.

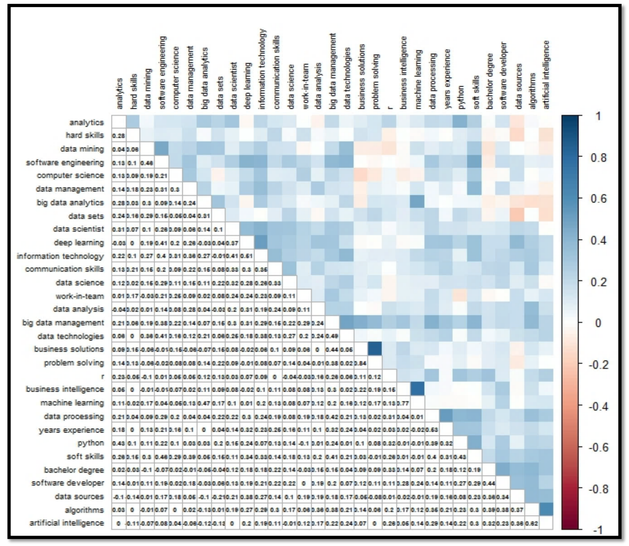

Figure 6: Social Network Analysis with undirected tie method.

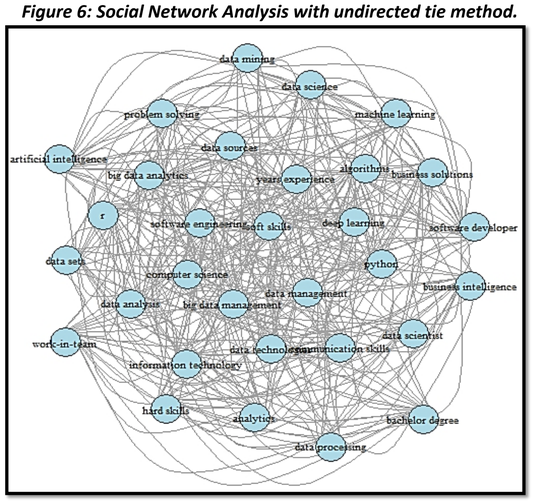

Figure 7: Social Network Analysis with directed tie method.