Vol. 14(2), pp. 53-64, February, 2020
DOI: 10.5897/AJBM2019.8885
Article Number: 03FC69A62893
ISSN: 1993-8233

http://www.academicjournals.org/AJBM

ACADEMIC JOURNALS
expand your knowledge

African Journal of Business Management

Full Length Research Paper

How universities fill the talent gap: The data scientist in the Italian case

Maddalena Della Volpe[1*] and Francesca Esposito[2]

[1]Department of Business, Management and Innovation System (DISA-MIS), University of Salerno, Italy.
[2]Department of Political and Communication Sciences, University of Salerno, Italy.

Received 13 September, 2019; Accepted 22 January, 2020

This research paper explores Italian study programs in data science in order to verify if knowledge and skills developed during the universities' path are fit with data scientist job demand. The issue is introduced considering the companies' growing need to derive insights from data, and consequently, to search for a staff with analytical expertise, the so-called data scientists quite rare. Literature review is focused on the data scientist's specific characteristics. According to the ideal profile, the data scientist should possess skills enabling the scientific collection, analysis and use of quantitative data in addition to managerial and communication skills, ensuring profitable interactions with decision-makers. The methodology introduces an innovative semi-automatic linguistic analysis of textual data, which enriches traditional statistical methods in text annotation and increasingly constitutes a key step to retrieve more and more precise information from large corpora. As results, the data scientist education in Italy is not widespread and the skills match highlights significant gaps between universities and companies in developing programming and software development skills. In conclusion, an intensive university-business cooperation in order to prepare future professionals, in line with technological trends and company requirements, could contribute to fill this gap, producing positive effects for the social and economic development.

Key words: Data scientist, business management, higher education, university, innovative skills, text mining, NooJ.

INTRODUCTION

The big data revolution (McAfee et al., 2012; Kitchin, 2014) has changed the way in which institutions, governments and companies rethink decision-making processes (Cukier and Mayer-Schoenberger, 2013). This has impacted on the job market and, as a result, several new professional figures are emerging to face the challenge posed by the considerable volume of data generated on the Web. Data scientists, analysts and engineers are in greater demand than ever before and, for many companies, they are still hard to find (Storey and Song, 2017). Thus, although in 2012 the Harvard Business Review baptised the data scientist as the sexiest job of the 21st century (Davenport and Patil, 2012), understanding and dealing with data remains a

*Corresponding author. E-mail: mdellavolpe@unisa.it.

major challenge for organizations (Schewe and Thalheim, 2008). The data scientist's profession is a complex one based on a multifarious set of capabilities and knowledge: according to the ideal profile, the data scientist should possess technical skills enabling the scientific collection, analysis and use of quantitative data in addition to managerial and communication skills ensuring profitable interactions with decision-makers and managers (Agasisti and Bowers, 2017). These skills can only be acquired by completing an interdisciplinary study path, but universities have been slow to meet this educational challenge: there is a fair number of data science degree programs, specializations and master's programs around the world (Dumbill et al., 2013), while undergraduate degree programs are scant and imprecise (Aasheim et al., 2015). What is needed is a strong response and a joint commitment from both educational institutions and companies in order to balance supply and demand in the field of data science (IBM, 2017).

As drivers of innovation, universities can influence the social and economic environment in which they operate (Goldstein, 2010; Kruss et al., 2012; Guerrero et al., 2016; Etzkowitz and Zhou, 2017). In an ever-changing world, therefore, universities must engage in upgrading strategies and policies in order to prepare students with knowledge, skills and aptitudes in line with technological trends and advances: the alignment of academic goals with the business world is thus essential in order to enhance the creation of future professionals (Perera et al., 2017).

Davenport and Patil (2012); Fisher et al. (2012); Granville (2014) and Besse and Laurent (2016) on the issue of data scientist tried to describe this professional figure, highlighting the characteristics and the main work tasks they carry out within companies: unfortunately, there is still no clear and shared definition, given the complex set of skills that a data scientist must have to be present in different market sectors.

The objective of this study is to explore the gap, in terms of developed skills, between universities' educational offer and the companies' competences requirements, explored by means of job demand on the business-networking website LinkedIn in Italy. The match detected refers to the wider debate about cooperation between universities and business in building programs, in order to allow graduates to acquire the right skills and mindset required by the job market.

In Italy, the Big Data Analytics market is growing steadily: in 2018, it was estimated to amount to 1.4 billion euros, +26% compared to 2017. According to Osservatori.Net (2018), investment in analytics, mostly in large companies, is focused on software development (45%), services (34%) and infrastructure (21%), and the demand for data science skills is increasing: 46% of large companies have already hired data scientists. Despite the growth of a ripe data science management model (from 17 to 31%), more than half (55%) of these apply a traditional organizational model. The adoption of descriptive analytics has resulted in an increase in the number of fast data initiatives, such as real-time advertising, fraud detection, predictive maintenance and new product development. The most serious obstacles to progress along this path are the lack of both skills and an internal workforce (53%), followed by scant management involvement (27%), as well as the difficulty in recruiting professionals with suitable skills (18%).

The data scientist's job requires a unique combination of skills, usually comprising both a solid foundation in data science and an innate talent for synthesizing complex ideas, thus addressing decision-making processes in organizations. The data scientists are so rare that some managers likely associate them with unicorns. Some researchers tried to give a definition of the data scientist by analysing the main work tasks in companies. On the contrary, other researchers qualify data scientists for their knowledge in statistics, computer science and information technology, in order to determine the appropriate skills to perform these certain tasks. We intend to contribute to this topic, sustaining a greater collaboration between universities and companies to fill the talent gap combining all the skills requested.

In order to understand what Italian universities are doing to fill the talent gap, we have collected data on current study programs in data science. The textual data retrieved on universities' official websites constitute the linguistic corpus, which we have processed using the Natural Language Processing (NLP) software environment NooJ. This tool gives researchers the opportunity to match large corpora with specific linguistic resources. The linguistic resources used in this paper are a set of automatically annotated local entries that express the data scientist's characteristic features, taken from the literature survey and from companies' job advertisements on LinkedIn. In this way, we explore the skills developed through data science education and match them with those most requested by companies in order to draw up guidelines by balancing supply and demand. After this introduction, we provide a survey of the literature on data science and then focus on data science education within specific study programs. This is followed by an explanation of the methodology used to process the textual data on study programs. The NLP experiment is then described and the main results are shown. The final section presents a discussion of the results and some concluding remarks.

THE DATA SCIENTIST: A LITERATURE REVIEW

In recent years, organizations' need to derive insights from data is growing, and consequently, companies are searching for a staff with analytical expertise. However, the so-called data scientists are quite rare.

Demand for data scientists started to rise in the late

20th century, especially among companies operating in the San Francisco area (Davenport, 2014). According to the Data Scientist Report 2018, data scientists are still in great demand and people with analytical expertise are being presented with new job opportunities several times a week. This survey involves 240 respondents from companies, 70% of whom still work with structured rather than unstructured data (Figure Eight, 2018). As a result, the advent of big data poses a number of challenges for management (Schewe and Thalheim, 2008; Chen et al., 2013), including difficulties in recruiting data scientists (Davenport and Patil, 2012). Hence, while many large companies, like LinkedIn, IBM, Macy's and General Electric are augmenting their teams with data scientists capable of managing big data technologies (Davenport and Dyché, 2013), other companies still seem to be struggling to source these rare professionals (Storey and Song, 2017).

Provost and Fawcett (2013) observe that, in order to serve the business demand for data scientists, it is important first of all to define the discipline of *data science*, pointing out its fundamental features and their relationship with other knowledge fields. The authors suggest that there are two reasons why this concept is often confused: firstly because it has been developed simultaneously with other concepts such as big data or data-driven decision-making; and secondly because the absence of academic programs leads people to associate the data science field only to what practitioners do, without identifying appropriate theoretical aspects.

From our point of view, Davenport and Patil (2012) provide one of the first definition of the data scientist: 'a high-ranking professional with the training and curiosity to make discoveries in the world of big data' (p. 72). However, there is a clear research gap in the formal definition of the 21st century's most prominent job (De Mauro et al., 2018; Hu et al., 2018), partly in view of the great number of different job roles that have been erased with the advent of big data (Miller, 2014), such as the data analyst, the data engineer, the big data expert or the big data architect. Giaume (2017) highlights how the data scientist is not a novelty and suggests that the innovative element of current data scientists probably lies in the fact that they now deal not only with numbers, but also with different type of contents, like images, audio and video. On the other hand, Besse and Laurent (2016) state that the new role of the data scientist should associate two types of approaches or logics: the *statistician logic*, which infers or checks for errors or risks in specific procedures, and the *computer scientist logic,* which designs the best strategy to minimize errors and optimize complex models in order to reach research objectives. The challenge for data scientists is methodological: innovative technological change affects choices of analysis strategy.

Some studies have attempted to classify data scientists according to their features. Granville (2014) identifies Vertical and Horizontal data scientists. *Vertical data scientists* have highly developed technical knowledge and skills. They are experts in statistics, computer science and operational research or hold an MBA, but they do not manage all these aspects together. *Horizontal data scientists*, on the other hand, combine vision with innovative data-driven techniques applied to unstructured data: they need to possess cross-disciplinary knowledge, including computer science, statistics, machine learning and domain expertise. These features can be related to the model by Davenport (2014), who recognizes five data scientist personalities: *the hacker*, who has the ability to write and program codes; *the scientist*, who can take decisions and possesses improvisation, impatience and orientation to the action; *the adviser*, who possesses strong communication skills; *the quantitative analyst*, who is able to use innovative quantitative techniques in statistics; and *the business expert*, who has a thorough knowledge of the business domain in which he/she operates.

Other authors have also defined the data scientist by underlining different aspects. For instance, Van der Aalst (2014) considers data scientists as the engineers of the future. He affirms that data science involves social sciences, industrial engineering and visualization: the data scientist is an engineer with quantitative, technical, creative, communicative skills, 'able to realize end-to-end solutions' (p.10). Dahr (2012) points out how machine learning skills are fast becoming a necessary skill set comprising statistics, computer science and problem-solving. In addition, text mining and knowledge of mark-up languages are also becoming fundamental.

A few studies focus on the typical data scientist workflow in companies. For instance, Fisher et al. (2012) interviewed 16 data analysts at Microsoft and identified their typical activities as consisting of acquiring data, choosing architecture, shaping the data to the architecture, writing and editing code, reflecting and iterating on the results. In a prior study, Kim et al. (2016) also interviewed 16 Microsoft data scientists and, two years later, presented a large-scale survey with 793 data scientists, again at Microsoft, in order to understand their educational background, the main work topics, the tools used and activities accomplished (Kim et al., 2018). Harris et al. (2013) investigate how 250 data scientists view their own skills, careers and experiences. The authors recognize four types of data scientist: Data Businesspeople, Data Creatives, Data Developers and Data Researchers, each of them with a profound expertise in a single skill set. They include the skill list obtained in the 'T-shape' model, made by Business, Machine Learning/Big Data, Mathematics/Operations Research, Programming and Statistics. Finally, Kandel et al. (2012) also interviewed 35 analysts in commercial companies ranging from healthcare to retail, finance and social networking. The authors recognize that analysts must have an aptitude for *discovery*, *wrangling*, *profiling*, *modelling* and *reporting*. Lastly, the role of visualization

skills is emphasized as an outcome of the whole data scientist workflow.

Universities' commitment in filling the talent gap

The heated debate over the skills data scientists need to have brings the role of education into play. According to Song and Zhu (2016), education is the key to success in data science: appropriate strategies are needed to prepare future professionals and universities should fast-track in creating degree programs for students (Fisher et al., 2012; Miller, 2014). According to Deloitte (2016), universities and colleges do not produce data scientists fast enough to meet business demands: despite the surge in data science programs the response of higher education seems to be inadequate because of the multifarious skills set that a data scientist should possess.

Some authors provide a description of current data science programs in order to assist universities in designing and developing undergraduate courses (Anderson et al., 2014; Aasheim et al., 2015; Baumer, 2015; Hardin et al., 2015; Asamoah et al., 2017). De Veaux et al. (2017) introduce an integrated curriculum that combines three disciplines offered separately in traditional courses like mathematics, computer science and statistics. Moreover, six main subject areas are identified in order to create a successful data science program: Data description and curation, Mathematical foundations, Computational thinking, Statistical thinking, Data modelling, Communication reproducibility and ethics.

In some cases, researchers pay attention to current or future programs that regard graduate students as analytics managers. Wilder and Ozgur (2015) propose an innovative Business Analytics program incorporating features of management and computer science in which they enhance the role of quantitative and technical skills together with soft skills like problem solving, teamwork and communication. Gupta et al. (2015) aim to develop a new training model, in which appropriate elective courses are added to existing curricula in order to foster the development of Business Intelligence skills, knowledge and experience for undergraduates, masters and MBAs. Finally, De Mauro et al. (2018) provide a big data skills framework that is useful for human resource recruiters and education providers in meeting the demand for four job families: business analysts, data scientists, developers, system managers. In particular, data scientists need knowledge of business impact, project management, database management and analytics.

In short, the literature reviewed shows a growing interest in data scientists and the majority of authors agree on the lack of academic programs. A stronger commitment to university-business collaboration is needed in order to design innovative educational paths and offer a better training. In the light of the literature examined, our proposed classification groups data scientist skills in nine clusters:

(i) Analytical skills. The ability to collect, analyse and interpret data, in order to help solve business problems and assist in decision-making.
(ii) Educational requirements. The degree of education, certifications, qualifications and the scientific domain of specialization.
(iii) Effective communication. The ability to communicate work results in one's own language but also in at least one foreign language (both in writing and orally).
(iv) Machine learning. The ability to research and develop algorithms and the capacity for automatic application of complex mathematical calculations to data.
(v) Knowledge management. Knowledge of the employer company and the ability to understand how the data can be useful in different business units.
(vi) Mathematics and Statistics. The ability to apply statistical models, concepts and processes to given situations using innovative tools.
(vii) Programming and Software development. The ability to write and program codes together with an understanding of big data architectures and infrastructures.
(viii) Soft skills. Interpersonal qualities and individual aptitudes, abilities such as teamwork and problem-solving, communication and self-management, critical thinking and ethics.
(ix) Visualization skills. The ability to turn data into innovative graphics or charts so as to uncover patterns, correlations and trends that will help people to understand which insights should be gained from the results.

We provide keywords or expressions used to describe the desirable features of a data scientist in Table 1, which will be used to build the linguistic resources necessary to carry out our text mining experiment.

MATERIALS AND METHODS

To compare the skills that universities intend to develop with those that companies seek to achieve their strategic goals, we combine quantitative and qualitative techniques in analysing textual data: the qualitative analysis in used to explore the meaning of words, while the quantitative analysis gives us the dimension of phenomenon in numerical terms. In business studies, this combination of quantitative and qualitative methods involves a sort of sensitivity to context, thus getting richer descriptions rather than only quantifiable metrics. When both approaches are used together, we can retrieve wider information.

Therefore, the analysis that we apply to textual data is not based only on key words, but it examines the context in which they are expressed and determines the meaning. We realised a semi-automatic textual analysis of the two main sources of information: on the one hand, all the texts referred to study programs and educational objectives; on the other hand, all the texts referred job advertisements that companies publish when they research the

Table 1. Words or expressions contained in each data scientist skill cluster.

Skill cluster	Some words or expressions
Analytical skills	Information retrieval; data storage; cloud computing; business intelligence; sentiment analysis; text mining; predictive analysis
Educational requirements	Bachelor's Degree; Master's degree; Doctor of Philosophy; PhD; MBA; Degree in Computer Science, Statistics, Mathematics, Physics
Effective communication	Communication ability; interpersonal communication; public presentations; ability to summarize; fluent in English; native speaker
Machine learning	Artificial Intelligence; neural networks; deep learning; learning algorithms; robotics; machine translation; pattern recognition
Knowledge management	Decision support systems; digital transformation; project management; market researches; customer satisfaction; business strategy
Mathematics and Statistics	Mathematical background; linear algebra; calculus; probability; statistical modelling; multi-dimensional data analysis
Programming and Software development	Solutions architecture; Python; Java; NoSQL; Web Development; Agile methodologies; real-time processing
Soft skills	Team working; problem-solving; leadership; goal orientation; emotional intelligence; motivation; entrepreneurial mindset; creativity; sense of responsibility
Visualization skills	Visualization of networks; interactive dashboards; reports; data visualization; reporting and analysis; dash-boarding

data scientist on LikendIn. We took into consideration the population of all Italian universities and all the announcements for the data scientist on the business-networking website LinkedIn, all published in the time when we started our data collection, in March 2019.

The Lexicon-Grammar: A conceptual framework for textual analysis

The investigation carrying out in this paper performes a combination of statistical and computational linguistics techniques of analysis. We introduce a semi-automatic linguistic analysis of textual data that enriches traditional statistical methods in text annotation and increasingly constitutes a key step to retrieve more and more precise information from large corpora. To understand better the method and the techniques used, we start from the Lexicon-Grammar (LG) conceptual framework, according to which, in a text, the elementary unit of meaning is the simple sentence. It is not possible taking in consideration words without considering the linguistic context in which they are expressed. Maurice Gross defines LG as the method and the formal description of the natural language in the second half of the 1960s. The LG involves the systematic syntactic description of the lexicon and goes beyond the exclusive search for general syntactic rules, regardless of the lexical material (Gross, 1968, 1979). In order to ensure good reproducibility of the observations, the affirmative simple sentence is considered as the minimal linguistic unit: a word acquires a certain meaning only in a certain syntactic context (Vietri, 2004). In other words, in order to derive high-quality information from texts, we cannot only consider single words and how many times they occur in the text, but we must observe the context in which words are inserted. In the LG methodology, therefore, the collection and analysis of a large quantity of linguistic resources and their continuous comparison with the reality of linguistic usages are crucial. This linguistic approach applied to NLP makes it possible to enrich statistical methods used in text annotation and increasingly constitutes a key step in the analysis of large corpora.

The LG uses electronic dictionaries (Appendix) to have definitions about special terms), lexicon-grammatical tables and local grammars to formalize the natural language. For this experiment, we have built a specific local grammar to define the data scientist semantic field. Local grammars are algorithms that, through syntactic, morphological and lexical instructions, formalize linguistic phenomena and enable the automatic processing of texts. By *local* we mean that these grammars can be used for the description of specific linguistic phenomena.

For instance, the data scientist grammar that we have created is a collection of 276 entries referring to data scientist features: 105 are single words, while 171 are multiword expressions. *Single words* are words without breaks or blanks, whereas *multiword expressions* are sequences of two of more simple words separated by a blank and characterized by a semantic atomicity (Gross, 1986). These expressions unambiguously identify a specific concept and, for this reason, multiword expressions are more relevant although less frequent than single words in terminology or technical lexicons. The linguistic resources are in Italian and English, due to the massive presence of English words in the data science semantic field. The local grammar contains ten graphs with the skill clusters presented in Table 1 above: there is one main graph and nine embedded ones. The words and expressions declared therein are examples of words inserted into each graph.

Collection, pre-processing and processing corpus

The first research phase consisted of a text collection. We collected texts from all the official websites of Italian universities offering study programs in data science for the academic year 2018-2019. In March 2019, we collected by means a scraping procedure published texts about program presentation, objectives, study plans, lessons, requirements, contents and learning methodology. In statistical terms, our data represent the whole population with reference to textual online materials about study programs in data science offered by all Italian universities. We gathered texts according to their educational degree: Bachelor's degree (BA), Master's Degree (MA) and Master postgraduate programs. In order to obtain a general overview of data science education in Italy, we also collected other types of information in a dataset that we would consider in the descriptive statistical analysis. The dataset, built

Table 2. Distribution of local entries in the corpus parts

Part	Size	AF	EF	SS
Bachelor's degree (BA)	8,235	64	37.05097	1.219891
Master's degree (MA)	70,854	319	318.7868	0.00965
Masters post graduate	33,153	122	149.1622	-1.22954

likewise in April 2019, contains the following fields taken from the official websites of Italian universities: program type, program name, degree class, program duration, university, location, department, internship, language, costs and requirements.

The second phase involved the cleaning and normalization of the retrieved texts so as to give them a format suitable for the corpus-processing phase. Linguists perform the text cleaning and normalization manually: for instance, white spaces and punctuation are removed in this process. Even numbers, dates, acronyms and abbreviations are *non-standard words* (Sproat et al., 2001) and had to be standardised in order to obtain the correct results in the linguistic analysis. Moreover, some words needed to be transformed from upper case to lower case, e.g. 'Big Data' to 'big data', and punctuation marks and some type of accents had to be deleted.

In the third phase, we included and processed all the texts in a single corpus using NooJ: a NLP software environment and corpus processor constructed by Silberztein (2003, 2015). The NooJ software environment makes it possible to process large corpora and develop orthographical and morphological grammars, dictionaries of simple or multiword expressions, as well as local and structural syntactic grammars. In other words, NooJ gives researchers the possibility to create specific linguistic resources and to store them in large corpora in the form of concordances, which represent the context in which the words are inserted.

RESULTS

Descriptive analysis

Data science is not widely taught in the Italian education system. Currently, out of 97 Italian universities (MIUR, 2019), 17 offer study programs in data science and big data analytics. In total, there are 21 study programs: 13 in Northern Italy, 6 in Central Italy and 2 in the South. As far as their educational level is concerned, 5% are BAs, 67% are MAs, 19% are first level Masters (after BA) and 9% are second level Masters (after MA). The degree classes to which the study programs belong are Statistics (9.5%), Physics (9.5%), Mathematics (9.5%), Computer Science (19%), Electronic Engineering (4.8%), Computer Science Engineering (4.8%), Techniques and Methods for the Information Society (14.3%). However, we can observe a prevalence of 2-year study programs (81%).

All the study programs identified are conducted in English. Indeed, the most common entry requirement (86%) is at least a B2 level of English, although only 14% of these need a formal certificate, while 57% of the study programs require a qualification in specific disciplines like Computer Science, Computer Engineering, Physics, Mathematics or Statistics. The cost of registering varies from 0 to 16,000 euros per annum, depending on the students' income, type of course and university. In particular, the BA in data science costs from 0 to 2,500 euros and the MA in Data Science costs from 1,350 to 14,000 euros; the cost of the first level Master's course varies between 3,200 and 6,000 euros, whereas the second level Master ranges from 9,000 to 16,000 euros. Finally, an internship is envisaged in the majority of study programs (81%), in 14% it is not provided for and in 5% no internship is specified.

Concordances and statistical analysis

According to the linguistic analysis within NooJ, the corpus consists of 112,242 characters and 18,478 tokens, of which 15,720 are word forms. In the phase of lexical acquisition by the corpus, we extracted a further 63 lexical entries to add to the data scientist local grammar: 60 multiword expressions and 3 single words. At present, there is a total of 399 lexical entries in the grammar, of which 108 are single words and 231 are multiword expressions. For example: *Matematica,* Noun (N), 'Mathematics'; *trattamento di dati*, Noun+ determiner+Noun (NdN), 'data processing'; *statistica inferenziale*, Noun+Adjective (NA), 'inferential statistics'; *scientific communication*, Adjective+Noun (AN); *finance for big data,* Noun+Preposition+Adj+Noun (NPAN).

At this point, we performed concordance analysis, which matches the local resources with word forms in the corpus. We found 505 matches in the corpus, distributed in word forms as follows: 21% N, 58% AN, 6% NA, 8% NdN, 2% NPN and 5% Acronyms (Acr.). The number of multiword expressions is higher than single words, which makes it possible to disambiguate a larger number of lexical entries and enables a deeper analysis of the meaning hidden in the corpus. The value of the Normalized Standard Deviation (NSD), indicating the distribution of words in relation to the number of matches in the corpus, is 0.0437. This point to a fairly regular distribution of lexical entries among all the texts. Table 2 reports the difference between corpus parts in terms of words frequency. The size of the texts (number of characters) is indicated in the first column, AF indicates the Absolute Frequency of matches in each text, EF is the Expected Frequency in each text and SS is the Standard Score (AF - EF) / SD, where SD is the Standard Deviation. In the MA, the AF and the EF values are

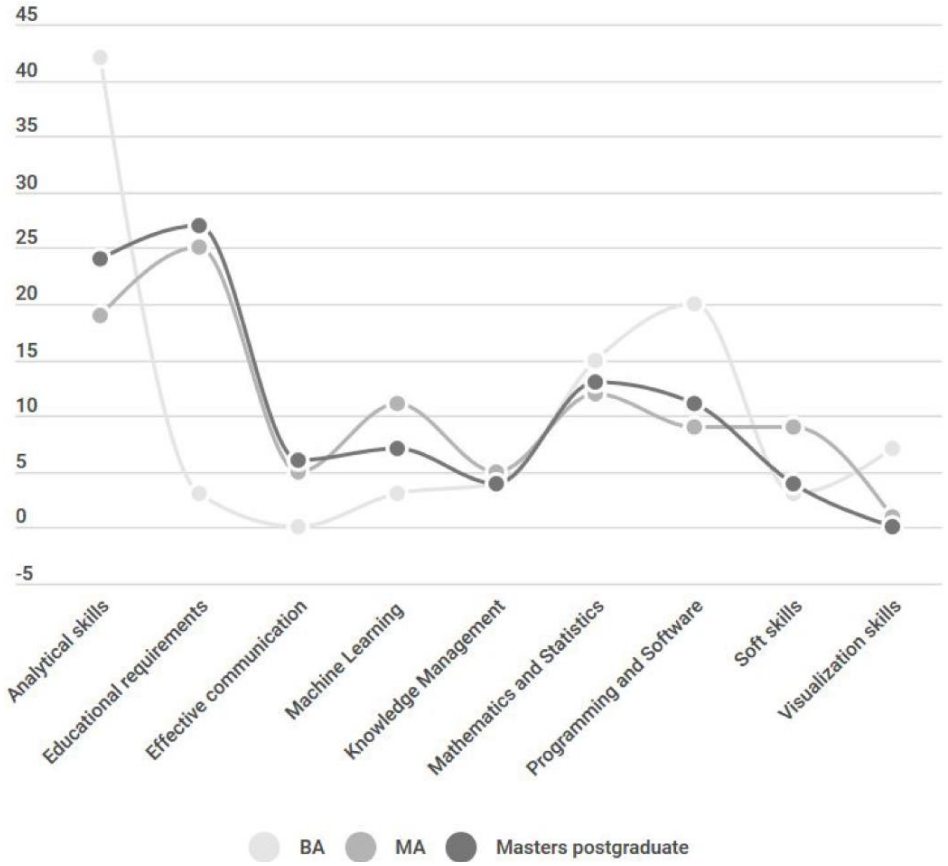

Figure 1. Skills cluster distribution in different study programs.

similar, so the value of SS is close to 0. This does not occur in the BA and Masters.

In order to ensure a clearer picture of word frequency in the corpus, we have put words with similar or identical meanings together. In the pre-processing phase, every possible variation of natural language has to be considered: after concordance analysis, lexical entries are regrouped according to different criteria. For instance, we combine the translation from Italian to English and vice versa, such as 'Game theory' and 'Teoria dei giochi' or even 'trattamento di dati' and 'data processing'; plural and singular entries, such as 'report' and 'reports'. Another criterion is referred to the lexical entries with the same meaning in terms of owned skills: 'data warehousing' with 'data warehouse', where the first identifies the storing activity and the second represents the storage, but in both cases, they refer to the same ability to cope with a data repository. Another issue in the NooJ processing phase concerned the difference between capital and small letters, which in some cases is essential to disambiguate entries; in other cases, it doubles word frequency (e.g. 'data processing' and 'DATA PROCESSING' or 'Data mining' and 'data mining'). We regrouped these lexical entries as far as possible. We were then able to identify the ten most frequent entries in the corpus of data science study programs in Italy: 'machine learning' (33) and 'statistics' (33), followed by 'Computer Science' (23), 'algorithmic methods' (18), 'data management' (18), 'data analytics' (18), 'data mining' (15), 'Engineering' (13), 'English' (13), 'Physics' (13). Concordance analysis highlights word distribution according to the nine clusters mentioned above. In the pre-processing phase, we divided lexical entries into nine hidden hubs composing the data scientist local grammar. The most relevant cluster was Analytical skills (24%), followed by Educational requirements (23%), Mathematics and Statistics (13%), Programming and Software development (11%), Machine Learning (10%), Soft skills (7%), Effective Communication (5%), Knowledge management (5%) and Visualization Skills (2%). Subdividing the findings into the BA, MA and Masters programs identifies the differences in cluster distribution. In other words, we learn what kind of skills the university aims to develop in students, according to the educational level (Figure 1). The Analytical skills cluster (42%) prevails in the BA, followed by Programming and Software (20%), Mathematics and Statistics (16%). The educational requirements cluster (MA= 25%; Masters= 27.9%) occurs more frequently in both MA and Masters, followed by Analytical skills (MA = 20%;

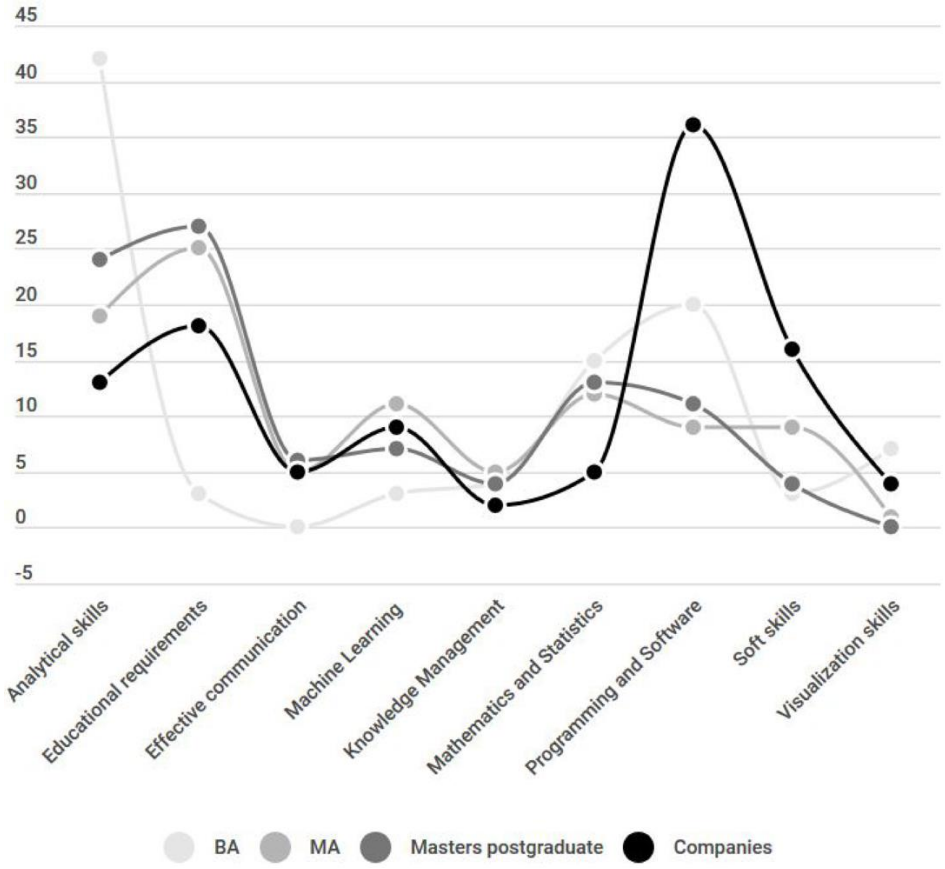

Figure 2. Comparison between companies' skill clusters and universities' skill clusters.

Masters= 12%) and, finally, Mathematics and Statistics (MA= 25%; Masters= 13%). The least important clusters are Effective communication for the BA and Visualization skills for the MA and the Masters. Another difference worth noting concerns the Programming and Software cluster, which is more important in the BA than in the MA and the Masters programs.

In order to verify the talent gap, a comparison has been made between what universities offer and what companies demand regarding data scientists' skills. In previous research (name deleted to maintain the integrity of the review process), we processed a corpus of 394 job advertisements on networking website LinkedIn so as to classify the skills required to be hired as a data scientist by companies in Italy. We applied the results of the linguistic analysis of companies' job advertisements to the universities study programs (Figure 2). The curve generated by the clusters distribution in companies appears to have a similar trend for the MA and Masters programs, except for the Programming and Software cluster which presents a lower value in the study programs compared with the companies' value. However, the MA and Masters programs appear to offer a better fit for the working world. The BA follows a different evolution, giving greater importance to Programming and Software.

DISCUSSION

Despite the increasing need that companies in Italy have for data scientists, higher education has been slow to react: in the 2018/2019 academic year, we find just 17 study programs. Moreover, they are not evenly located throughout the country, because their concentration has been relived mainly in Northern Italy, the most highly industrialized part of the country. The study programs offered by universities are not only few, but most of them, are specialized post-degree paths lasting one or two years, with just one path lasting three years. This kind of study path is too short to acquire the adequate and complex skills set necessary to become a data scientist. As we explored in literature, this profession involves not only technical skills enabling the scientific collection, analysis and use of quantitative data, but also a multifaceted set of competences, being able to predict problems and to communicate the best solutions for the company according to Agasisti and Bowers (2017) and Van der Aalst (2014). It means also being able to handle the whole process of storage, analysis, integration,

visualization and communication of data. The data science profession supersedes the traditional profession of the data scientist (Davenport, 2014). While once it was sufficient to acquire technical knowledge in order to become an expert in statistics, computer science or operational research, nowadays the data scientist needs to develop a cross-disciplinary knowledge, which includes traditional domains and adds a more innovative and data-driven vision. As far as higher education is concerned, this calls for a commitment in designing new study paths for data scientists, while not neglecting soft skills, creativity and data visualization. In a changing environment, soft skills assume greater importance in increasing students' responsibility and their ability to adapt to a dynamic job market. It entails understanding the commitment and specific characteristics of the workplace: being ethical, balanced, able to work autonomously or as part of a team, leadership, results orientation, management and continuous improvement skills (Succi and Canovi, 2019). Differently from suggestions by literature analysed and companies requirements, Italian study programs are not focused on enhancing soft skills because they are only recognised as having a low value. Our linguistic analysis associates them exclusively with words like 'problem solving', 'active participation' and 'team working', reducing the richness and the variety of the concept which includes the ability to plan and communicate work results, self-management, critical thinking and creativity.

The talent gap, which emerges in a very visible way from the study, underlines the increasingly necessary cooperation not only in the phase of leaving the university, but also within the training path. For too long universities have been disconnected from the companies, producing a marked misalignment between job supply and demand, leaving both unsatisfied the employment needs of companies and the professional aspirations of graduates.

Within companies, the conflict between chaos and discipline, order and disorganization, strictness and imagination is mostly unresolved: overall, companies prefer regulation to imagination or creativity. However, creativity helps to add innovation to the organization, because it breaks the rules and makes it possible to assimilate new elements. As a result, the data scientist profile needs to open up to creativity, so as to be able to handle numbers with an innovative vision. Data visualization is likewise undervalued in the study programs considered. Although data mining allows us to extract knowledge, there are many hidden or unknown aspects, which can be analysed or identified only with human judgement: data visualization 'may reveal patterns that would otherwise remain unnoticed' (Van der Aalst, 2014: 12) by exploiting human capabilities for perception.

Our findings reveal further information on the distribution of skill clusters in study programs: the only BA study program is mostly focused on the development of Analytical skills such as big data analytics, data integration and data warehousing: concepts aiming to introduce data analysis. The MA and Masters are more likely in Educational requirements: these study programs are oriented towards deep knowledge fields that have already been studied in a previous degree path, such as Statistics, Mathematics, Physics and Computer Science, while effective communication, knowledge management, soft skills and visualization are largely undervalued.

Finally, the comparison between companies and universities reveals some difference between what education offers and what business requires. Companies need a strong specialization in Programming and Software, which is less relevant in universities' study programs. From the companies' viewpoint, data scientists must be able to use programming languages that allow them to access, explore and model data; data scientists must be skilled in software development, such as Java and C++, and familiar with many aspects of computational science and software engineering.

An internship is envisaged in most of the study programs analysed, in recognition of the importance of interaction between university and industry, but it is generally too short to enable this practical experience to generate interesting results in terms of new knowledge. We look forward to the internship being improved in terms of duration and quality (Della Volpe, 2017).

The challenge for higher education is to design new interdisciplinary paths, integrating knowledge from STEM disciplines (Science, Technology, Engineering, Maths), with those from humanities in an innovative vision. While the university programs considered are entirely focused on STEM disciplines, companies are already looking for creativity and communication skills. This means that the future is moving towards new frontiers and universities should not neglect this.

At the same time, it is important to remember that as data dissemination comes from a wide variety of data sources, it requires a rapid shift in the workforce in order to meet demand in the job market. In order to fill the gap between the academic and the business worlds, a new up-skilled and re-skilled workforce needs to be developed. As technologies change rapidly, students and workers should be prepared to engage in continuous training and learning in order to be employed in future jobs, the ones that have not yet emerged. The next generation of students should be trained in data literacy so as to manage data relevance while they are still at school (Deloitte, 2019). Universities and industries could collaborate by creating common data labs or communities so to work together in solving real business problems.

Conclusion

The challenge for universities is to integrate new study

programs in more interdisciplinary paths; the challenge for enterprises give space to the creativity, in order to view common collaboration. We limited our analysis to Italian case: for a comprehensive and worldwide overview, one should analyse other context in different countries universities. At the same time, we took into consideration only the public information provided by official universities' websites, because they constitute the best way to know the educational offer. However, we actually ignore what universities communicate on other social platforms, such as Twitter, Facebook, Instagram, LinkedIn. Besides, we are aware that this data provided by the universities' official website are related to the academic year 2018-2019: they can change year by year and it should be realized a longitudinal analysis to catch the change. Future researches could focus on comparative studies on this topic to highlight similarities and differences among countries.

CONFLICT OF INTERESTS

The authors have not declared any conflict of interests.

REFERENCES

Aasheim CL, Williams S, Rutner P, Gardiner A (2015). Data analytics vs. data science: A study of similarities and differences in undergraduate programs based on course descriptions. Journal of Information Systems Education 26(2):103-115.

Agasisti T, Bowers AJ (2017). Data analytics and decision making in education: towards the educational data Scientist as a key actor in schools and higher education institutions. Handbook of Contemporary Education Economics pp. 184-210. Available at: https://doi.org/10.7916/D8PR95T2

Anderson P, Bowring J, McCauley R, Pothering G, Starr C (2014, March). An undergraduate degree in data science: curriculum and a decade of implementation experience. In: Proceedings of the 45th ACM technical symposium on Computer science education pp. 145-150. Available at: https://doi.org/10.1145/2538862.2538936

Asamoah DA, Sharda R, Hassan Zadeh A, Kalgotra P (2017). Preparing a data scientist: A pedagogic experience in designing a big data analytics course. Decision Sciences Journal of Innovative Education 15(2):161-190.

Baumer B (2015). A data science course for undergraduates: Thinking with data. The American Statistician 69(4):334-342.

Besse P, Laurent B (2016). De Statisticien à Data Scientist. Statistique et Enseignement. Société Française de Statistique 7(1):75-93.

Chen J, Chen Y, Du X, Li C, Lu J, Zhao S, Zhou X (2013). Big data challenge: a data management perspective. Frontiers of Computer Science 7(2):157-164.

Cukier K, Mayer-Schoenberger V (2013). The rise of big data: How it's changing the way we think about the world. Foreign Affairs 92:28.

Dahr V (2012). Data science and prediction. CeDER Working Papers, CeDER-12-01. Available at:https://archive.nyu.edu/handle/2451/31553

Davenport T (2014). Big data at work: dispelling the myths, uncovering the opportunities. Harvard Business Review Press.

Davenport TH, Patil DJ (2012). Data scientist. Harvard Business Review 90(5):70-76.

Davenport TH, Dyché J (2013). Big data in big companies. International Institute for Analytics, 3. Available at: https://docs.media.bitpipe.com/io_10x/io_102267/item_725049/Big-Data-in-Big-Companies.pdf

Della Volpe M (2017). Assessment of internship effectiveness in South Italy Universities. Education+ Training 59 (7/8):797-810.

Deloitte (2016). Analytics trends: The next evolution. Report 2016. Available at: https://www2.deloitte.com/content/dam/Deloitte/ca/Documents/deloitte-analytics/ca-EN-16-4454H%20Analytics%20Trends_AODA.pdf

Deloitte (2019). Leading the social enterprise: Reinvent with a human focus. Global Human Capital Trends. Available at:https://www2.deloitte.com/content/dam/insights/us/articles/5136_HC-Trends-2019/DI_HC-Trends-2019.pdf

De Mauro A, Greco M, Grimaldi M, Ritala P (2018). Human resources for Big Data professions: A systematic classification of job roles and required skill sets. Information Processing and Management 54(5):807-817.

De Veaux RD, Agarwal M, Averett M, Baumer BS, Bray A, Bressoud TC, Kim AY (2017). Curriculum guidelines for undergraduate programs in data science. Annual Review of Statistics and Its Application 4:15-30. https://doi.org/10.1146/annurev-statistics-060116-053930

Dumbill E, Liddy ED, Stanton J, Mueller K, Farnham S (2013). Educating the next generation of data scientists. Big Data 1(1):21-27.

Etzkowitz H, Zhou C (2017). The triple helix: University–industry–government innovation and entrepreneurship. London, UK: Routledge.

Figure Eight (2018). Data Scientist Report 2018. Available at https://visit.figure-eight.com/WC-2018-Data-Scientist-Report.html

Fisher D, DeLine R, Czerwinski M, Drucker S (2012). Interactions with big data analytics. Interactions 19(3):50-59.

Giaume A (2017). Data scientist: Tra competitività e innovazione. FrancoAngeli.

Goldstein HA (2010). The entrepreneurial turn and regional economic development mission of universities. The Annals of Regional Science 44(1):83.

Granville V (2014). Developing analytic talent: Becoming a data scientist. Hoboken, USA: John Wiley and Sons.

Gross M (1968). Grammaire transformationnelle du français. Syntaxe du verbe, 1. Paris, France: Larousse.

Gross M (1979). On the failure of generative grammar. Language 55(4):859-885.

Gross M (1986). Lexicon-grammar the representation of compound words. Paper presented at The 11th International Conference on Computational Linguistics. Available at:https://doi.org/10.3115/991365.991367

Guerrero M, Urbano D, Fayolle A, Klofsten M, Mian S (2016). Entrepreneurial universities: emerging models in the new social and economic landscape. Small Business Economics 47(3):551-563.

Gupta B, Goul M, Dinter B (2015). Business Intelligence and Big Data in Higher Education: Status of a Multi-Year Model Curriculum Development Effort for Business School Undergraduates, MS Graduates, and MBAs. Communications of the Association for Information Systems. CAIS 36:23.

Hardin R, Nicholas J, Horton D, Nolan B, Baumer O, Hall-Holt P, Murrell R, Peng P, Roback D, Temple Lang, Ward MD (2015). Data science in statistics curricula: Preparing students to 'think with data. The American Statistician 69(4):343-353.

Harris H, Murphy S, Vaisman M (2013). Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work. Newton, MA: O'Reilly Media, Inc.

Hu H, Luo Y, Wen Y, Ong YS, Zhang X (2018). How to Find a Perfect Data Scientist: A Distance-Metric Learning Approach. IEEE Access 6:60380-60395. Available at:https://doi.org/10.1109/ACCESS.2018.2870535

IBM (2017). The Quant Crunch. How the demand for data science skills is disrupting the job market. Available at:https://www.ibm.com/analytics/us/en/technology/data-science/quant-crunch.html

Kandel S, Paepcke A, Hellerstein JM, Heer J (2012). Enterprise data analysis and visualization: An interview study. IEEE Transactions on Visualization and Computer Graphics 12:2917-2926.

Kim M, Zimmermann T, DeLine R, Begel A (2016). The emerging role of data scientists on software development teams. ICSE '16 Proceedings of the 38th International Conference on Software Engineering pp. 96-107. Available at:

https://doi.org/10.1145/2884781.2884783
Kim M, Zimmermann T, DeLine R, Begel A (2018). Data scientists in software teams: State of the art and challenges. IEEE Transactions on Software Engineering 44(11):1024-1038.
Kitchin R (2014). The data revolution: Big data, open data, data infrastructures and their consequences. Newcastle, UK: Sage.
Kruss G, Visser M, Haupt G, Aphane M (2012). Academic interaction with external social partners: Investigating the contribution of universities to economic and social development. Education and skills development. Cape Town, South Africa: HSRC Press.
McAfee A, Brynjolfsson E, Davenport TH, Patil DJ, Barton D (2012). Big data: the management revolution. Harvard Business Review 90(10):60-68.
Miller S (2014). Collaborative approaches needed to close the big data skills gap. Journal of Organization Design 3(1):26-30. https://doi.org/10.7146/jod.9823
MIUR (2019). Istituzioni universitarie accreditate. Available at: https://miur.gov.it/istituzioni-universitarie-accreditate
Osservatori.Net (2018). Il mercato dei Big data in Italia. Osservatorio Big Data Analytics & Business Intelligence. Available at: http://www.osservatori.net
Perera S, Babatunde SO, Zhou L, Pearson J, Ekundayo D (2017). Competency mapping framework for regulating professionally oriented degree programmes in higher education. Studies in Higher Education 42(12):2316-2342.
Provost F, Fawcett T (2013). Data science and its relationship to big data and data-driven decision making. Big Data 1(1):51-59.
Schewe KD, Thalheim B (2008). Semantics in data and knowledge bases. International Workshop on Semantics in Data and Knowledge Bases, Berlin, Heidelberg: Springer pp. 1-25.
Silberztein M (2003). NooJ manual. Available at: http://www.nooj4nlp.net/NooJManual.pdf
Silberztein M (2015). La formalisation des langues: l'approche de NooJ. London, UK: ISTE Editions.
Song IY, Zhu Y (2016). Big data and data science: what should we teach? Expert Systems 33(4):364-373.
Sproat R, Black A, Chen S, Kumar S, Ostendorfk M, Richards C (2001). Normalization of non-standard words. Computer Speech and Language 15:287-333.
Storey VC, Song IY (2017). Big data technologies and management: What conceptual modeling can do? Data and Knowledge Engineering 108:50-67.
Succi C, Canovi M (2019). Soft skills to enhance graduate employability: comparing students and employers' perceptions. Studies in Higher Education 2019:1-14.
Van der Aalst WM (2014). Data scientist: The engineer of the future. Enterprise interoperability VI, Interoperability for Agility, Resilience and Plasticity of Collaborations. pp. 13-26. Available at:https://doi.org/10.1007/978-3-319-04948-9_2
Vietri S (2004). Lessico-grammatica dell'italiano. Metodi, descrizioni e applicazioni.
Wilder CR, Ozgur CO (2015). Business analytics curriculum for undergraduate majors. INFORMS Transactions on Education 15(2):180-187.

APPENDIX

Concordance: collection of all the co-texts of the same word (pivot) in the corpus.
Corpus: collection of texts, consisting of one or more elements (fragments), consistent with each other for study purposes; when the collection of texts that makes up the corpus is large (several tens, hundreds or thousands of fragments), it is possible to associate structured information (coded data constituting quantitative or qualitative variables) to each element of the collection.
Disambiguation: elimination of the ambiguity of a lexical unit; action to attribute to a word the right grammatical category or the authentic meaning in a given context.
Electronic dictionary: it is a dictionary stored in the form of computer data (lexical database) rather than in human readable format. An electronic dictionary can be loaded into a database and interrogated by means of special software. The term electronic dictionary is also used to refer to an electronic vocabulary or lexicon, such as those used by spelling checkers. If a dictionary is structured through a hierarchy of supertype-subtype concepts, it is called taxonomy. If it also contains other relationships between concepts, it is called ontology. Search engines use vocabularies, taxonomies or ontologies to optimize search results. Specialized electronic dictionaries are, for example, morphological or syntactic dictionaries.
Entry: unit of a list or of a dictionary or of other lexical index.
Lemma: pair of information [canonical form, grammatical category] present in a language dictionary.
Lexical Analysis: level of study of the language in a corpus; the domain of Lexical Analysis is the vocabulary of the corpus; the product of a Lexical Analysis activity is the annotation of the lexical units.
Natural Language Processing (NLP): natural language treatment; set of automatic procedures for linguistic and / or semantic recognition of the words or of the sentences in a text.
Occurrence: (token, reply), each appearance of a word in the text; the frequency of a word in a text is given by the number of its occurrences (more properly we speak of normalized frequency).
Semantics: discipline of linguistics that studies the meaning of words or of sentences. The semantic units of analysis are lexemes, that is, the words studied from the point of view of meaning. The set of lexemes of a language constitutes its lexicon.
Text Mining (TM): exploration and "excavation" activity in a deposit of textual materials (corpus) for information retrieval and extraction; complex procedures for extracting knowledge, aimed at creating value, from vast documentary bases of companies or institutions
Text Normalization: Transforming non-standard words in a readable format for the computer. For example, white spaces and punctuation are removed in this process. The use of extensive social media has resulted is a new form of written text, this poses new challenges to natural language processing.
Text unit: lexical analysis unit for the automatic analysis of a text; occurrence; type; V = verb, N = noun, A = adjective, AVV = adverb etc.
Textual Analysis: level of study of the occurrences (single appearances of the lexical units) of a corpus; the domain of Textual Analysis is the set of fragments of the corpus; the product of a Textual Analysis activity is the annotation of the context units (categorization of fragments or documents).
Textual data: information about phenomena expressed by words. The textual sources are, therefore, interviews, open questions of a questionnaire, reports of a focus group, political speeches, material taken from the Internet, documents and much more.
Textual: attribute inherent in the text that is concerning the development of discourse in the corpus
Token: single occurrence or replica of a type; the set of tokens of a corpus expresses its extension or amplitude in occurrences.
Tokenization: process of segmenting the text into occurrences or tokens, based on a sequence of characters - defined as belonging to an alphabet - delimited by separators.
Vocabulary: list of the different lexical units of a corpus with the corresponding occurrences.
Word: conventional and generic term to identify the text analysis unit.