sustainability

MDPI

Article

Skill Needs for Early Career Researchers—A Text Mining Approach

Monica Mihaela Maer-Matei [1,2], Cristina Mocanu [1], Ana-Maria Zamfir [1,*] and Tiberiu Marian Georgescu [2]

[1] National Scientific Research Institute for Labour and Social Protection, 6-8 Povernei Street, 010643 Bucharest, Romania; matei.monicamihaela@gmail.com (M.M.M.-M.); mocanu@incsmps.ro (C.M.)

[2] Department of Economic Informatics and Cybernetics, The Bucharest University of Economic Studies, 6 Piata Romana, 010552 Bucharest, Romania; tiberiugeorgescu@ase.ro

* Correspondence: anazamfir2002@yahoo.com; Tel.: +40-21-3124069

Received: 9 April 2019; Accepted: 13 May 2019; Published: 15 May 2019

check for updates

Abstract: Research and development activities are one of the main drivers for progress, economic growth and wellbeing in many societies. This article proposes a text mining approach applied to a large amount of data extracted from job vacancies advertisements, aiming to shed light on the main skills and demands that characterize first stage research positions in Europe. Results show that data handling and processing skills are essential for early career researchers, irrespective of their research field. Also, as many analyzed first stage research positions are connected to universities, they include teaching activities to a great extent. Management of time, risks, projects, and resources plays an important part in the job requirements included in the analyzed advertisements. Such information is relevant not only for early career researchers who perform job selection taking into account the match of possessed skills with the required ones, but also for educational institutions that are responsible for skills development of the future R&D professionals.

Keywords: skills; researchers; early career; text mining

1. Introduction

Research and development activities are one of the main drivers for progress, economic growth and wellbeing in many societies. Even if the investments made in research have been questioned by some stakeholders, technological developments are needed for more science advancements in order to push forward sustainable economic development, especially in emerging economies aiming to catch up with developed countries [1]. In the European Union (EU), the main policy instrument in this field, the European Research Area (ERA), promotes several priorities such as effective national research systems based on investments and national competition, transnational cooperation and competition, open labour market for researchers, gender equality, and optimal circulation of knowledge. Thus, current policies promote a Europe based on the freedom of movement of people and knowledge. In order to achieve this goal in R&D, a number of initiatives have been promoted such as the European Charter for Researchers and Code of Conduct for their Recruitment. The key purpose is to support researchers' movement across borders, sectors and disciplines. Such ambitious goals are supported by the EURAXESS platform which is a pan-European initiative that provides information on job opportunities and supports research careers. This initiative is endorsed by the European Union, member states and associated countries. The number of job vacancies advertised on the EURAXESS platform witnessed an increase of 7.8% in 2012–2014, followed by a decline of 5% in 2015–2016 [2]. However, the share of researchers expressing their satisfaction with the level of openness, transparency and merit-basis of the recruitment processes increased by 7.5% in the 2015–2016 period [2]. Concluding, the EURAXESS platform plays a key role in supporting an open labour market for researchers. The information provided via EURAXESS takes the form of job vacancies and funding opportunities advertisements, aiming to allow for the match between the supply and demand for researchers. The development of researchers' careers and their movement across borders, sectors and disciplines are influenced by the attractiveness and quality of information that are available for them. The main goal of this article is to explore and reveal key dimensions that characterize the entry level research labour market for selected scientific fields.

According to the European Framework for Research Careers, four professional categories exist among researchers, irrespective of their working context (universities, research institutions, NGOs, companies): (1) First Stage Researcher, (2) Recognized Researcher, (3) Established Researcher and (4) Leading Researcher. This study is focused on the entry level research labour market, namely on positions for first stage researchers. Usually, first stage researchers are PhD candidates who carry out research activities under supervision, have a good knowledge of their field of study, are able to collect data under supervision, as well as to analyze and assess complex ideas and to present their research outcomes.

Previous studies on career management of R&D professionals have showed that individuals respond and take decisions based on the structure of available opportunities [3], meaning that they consider the perceived rewarded activities when they develop their career orientation and strategies in order to reach expected career outcomes [4,5]. On the other hand, following the idea that career decisions and challenges vary significantly by career stage [6], we focus our analysis on first stage research positions. Previous studies on R&D professionals found that, during the exploration stage, career goals include the understanding of personal abilities and interests, evaluation of job requirements, and integration within research teams. Entry-level researchers face the need to develop their professional identity, to contribute with their knowledge and competences within the organization and team, as well as to cope with challenging tasks [7,8].

As opposed to other sectors, career systems in R&D have been extensively influenced by issues related to the level of professional competences of the researchers and other relevant skills such as team work and problem solving and less by the traditional advancement in the organizational hierarchy [9–11]. Thus, attracting and retaining researchers with the right mix of knowledge and competences became a key factor for more and more organizations [12] as individuals make job selections that are consistent with their personal orientations and profile [13,14]. The volume of researchers represents an important input for the innovation processes [15–17]. Many scholars consider that it is important to better understand the reactions of researchers to various career opportunities [4,18–20]. Career choices are made on the basis of career orientation, which represents a mix of self-perceived preferences, talents, needs and values [18,21,22]. Five types of career orientation have been identified among R&D professionals: technical orientation, manager orientation, project orientation, technical transfer orientation and entrepreneur orientation [20]. However, different career orientations share many common competences, values and professional roles [23]. One common challenge is that R&D professionals face rapidly changing demands determined by new technologies developments [24].

The way individuals respond to various job opportunities is explained by the person–organization fit theory which refers to the way the profile and skills of workers match with the needs, practices and expectations of the organizations [25,26]. From this point of view, information on job opportunities that are provided by organizations to first stage researchers shape the way individuals make job selections in the R&D sector. While many studies analyze career orientation and choices of R&D professionals by exploring data collected from researchers [18,21,22], this article is focused on information coming from organizations in the form of job vacancies advertisements. Such information is relevant not only for early career researchers who perform job selection taking into account the match of possessed skills with the required ones, but also for educational institutions that are responsible for skills development of the future R&D professionals. Various innovative approaches have been developed in order to better inform education and training institutions with respect to the nature and level of skills required from their graduates [27,28]. This article proposes a text mining approach applied to a large amount of data extracted from job vacancies advertisements, aiming to shed light on the main skills and demands that characterize first stage research positions in Europe.

2. Skills for RDI Sector: Some Hints from the Literature

Although the purpose of identifying skills relevant to research and innovation might seem appealing for decision makers in the area of education and training, and although there are several research endeavors aiming to provide some hints, finding the links between skills and RDI and understanding the policy relevance of the results are not easy tasks.

One of the most well-known measures of skill needs is the required level of education. As the share of higher education graduates as well as the share of doctoral and postdoctoral graduates increased in the population, the minimum level of education required for entry level positions in RDI and universities increased to doctorate level. PhD holders are among the most mobile populations, the international mobility often starting from the training period/program [29,30], so a better knowledge on required skills could improve the PhD holders' mobility, as well as the knowledge flows among European countries. Current RDI strategies aim to support the increase of PhD holders' numbers for a specific theme/research sector, being considered that usually a PhD veils some mix of skills that supports research and innovation [29,31].

Apart from the apparent consensus on the minimum required level of education in RDI, findings from the scientific literature are very heterogeneous, as a lot of skills and personal characteristics were under scrutiny and proved to influence research and innovation ideas and outputs [29,32]. The RDI sector is a very heterogeneous one, and studies carried out in the field used different typologies and focused on the role of different skills, not to mention the different conceptual approaches of skills and innovation used. Although the meaning of skills varies a lot through the literature, we use for this paper a broad sense of the concept, covering abilities, competences, knowledge, as well as personal attributes [29,33,34].

Studies addressing skills for the RDI sector are rarely comparable across industry [32], addressing mainly the corporate side of the sector and usually finding a mix of skills supporting research and/or innovation. The mix of required skills covers basic skills, technical skills, academic or methodologic skills, soft skills, etc. [29,32].

The mix of skills needed in RDI varies along sectors (business, university, NGOs), according to industry structure and competitiveness, type of RDI (fundamental, empirical, etc.) or type of innovation. Higher sectoral skills lead to higher sectoral productivity [35], as well as to higher investments in R&D [36], so the sector's characteristics, its structure and competitiveness, could influence the required mix of skills. Methodological limits in introducing sectoral and specific skills in comparative surveys also limit the possibility to identify specific skills supporting research and innovation and urge for more in-depth studies at the level of sub-sectors and occupations. Skill needs in RDI usually imply both theoretical and practical skills [32].

Big innovations and outputs are more likely to be produced by highly specialized companies [37], so technical and methodological skills remain at the core of job requirements in RDI, while communication, teamwork, sharing, etc. increase their importance.

Leadership, management and entrepreneurial skills are also addressed by the scientific literature, but are treated on a rather separate track. Management and entrepreneurial skills can be considered as transversal skills along the entire RDI sector, increasing self-regulation and adaptability, irrespective of sector specificities, but also fostering and mentoring the organizational space where innovation might appear. Entrepreneurial skills foster spill overs and contribute to increasing R&D returns [38]. Managerial and leadership skills are crucial not only for better positioning the company/organization on the market, but also to develop cooperation with other stakeholders and competitors in the field [39].

Globalization, ICT and the increasing importance of green skills are among the drivers of change for the future skill demands of RDI sector [28]. Globalization and ICT are changing the way economies work, increasing competitiveness and urging for collaboration. Soft skills such as communication,

communication in foreign languages, teamwork, working in multicultural teams and organizations, and working in multidisciplinary/interdisciplinary teams might become more and more important. Apart for the increasing importance of the so-called soft skills, globalization in RDI also leads to increasing levels of specific and technical skill needs. Large amounts of data available due to internet development call for new methods and skills to collect, organize and analyze them. Globalization also urges for skills that can support comparable studies, both quantitative and qualitative. Mass education is less probable to provide such a high level of skills, so self-learning and learning to learn are among the skills underpinning skills development in RDI.

Sustainability-oriented innovation changes the way economies operate, new green skills, green occupations and even green sectors emerging. Also, more responsible and ethical attitudes towards environment, culture, and communities becomes mandatory in RDI, although their future impact on the RDI is hard to be estimated [29,40].

The current findings point to a broader set of skills needed in RDI, with different sub-sectors needing different mixes of skills in different contexts and for different purposes [29,32]. This is why pulling out a core set of skills to substantiate teaching and learning policies in the field might be a tricky task, asking in fact for stakeholders' involvement in curriculum design, as well as for more specific studies undertaken for different sectors and occupations in order to provide more detailed findings.

The mix of skills needed in RDI is also in line with the mix of technical, communication, IT, ethical, legal and data science skills needed to support the objective of promoting and developing Open Science [41].

3. Data Gathering Process

Text mining represents a solution to the research challenge induced by new data sources such as text data posted on the web. This has led to an increase in the amount of data that can be extracted and analyzed in different domains. For example, the content analysis of job advertisements is one research topic based on voluminous textual data providing findings about training needs or technical skills required for specific jobs. This information could be useful for academic institutions in updating their curricula or for individuals in their career planning or for job analysts. The studies developed for identifying the skills valued by employers using online job vacancies focused on information regarding the activities associated with the job and on the attributes required from applicants. Some of the researches focused on specific jobs: big data jobs [42], IT jobs [43], information systems [44], and librarians [45], meanwhile other studies analyzed the communalities encountered for different professions [46,47]. The studies highlighted that constantly a combination of technical and soft skills is required. Previous researches on data analytics positions in the business sector found a common set of soft and transferable skills such as decision making, organization, communication and structure data management [48]. The findings were rather limited with respect to a similar set of technical skills, only statistics and programing skills being mentioned as common to the scrutinized jobs [48].

Big volumes of data were required as a source for quantitative research. In the planning stage of the study presented in this article, data were collected manually via a web browser. When the authors comprehended the potential of the research, they decided to develop tailor-made solutions to collect data automatically.

The source of data was EURAXESS web platform. EURAXESS is a pan-European initiative backed by the European Union which provides valuable information and resources to researchers. For this article, the authors were interested in collecting information about jobs offers in the research field [49]. Data collected is from public pages and it is used for research purposes only.

The authors chose five domains to analyze: Computer Science, Economics, Engineering, Environmental Science and Mathematics. Data about research job offers were gathered for the 1 May to 27 October 2018 timeframe. Data collected included the research field, researcher profile, date, description and requirements. Many of the job listings include more than one research field.

If at least one of them was among the six mentioned above, the data about the page were gathered. The authors decided to keep only the job listings that were accessible to early career researchers. Therefore, the specifications of the researcher profile had to include First Stage Researcher (R1), which according to EURAXESS, includes "individuals doing research under supervision in industry, research institutes or universities", including PhD candidates, but not PhD holders [49]. Some job listings were dedicated to First Stage Researchers, others were open for more experienced researchers as well. Date, description and requirements were gathered for all the pages that respected the criteria described above.

The solution was developed in Python 3 programming language, using Scrapy—An open-source framework [50]. Previous studies [51,52] have used similar technologies to collect big volumes of data automatically in the absence of an API (Application Programming Interface). Article [53] describes in detail the process of data acquisitions, difficulties encountered and solutions to solve them.

Although the solution is able to extract data much faster, a download rate of 36 pages/minute was set. The authors were very careful not to affect EURAXESS server performances. The solution is designed to collect from every scraped page only the relevant data. That is possible by finding the CSS selectors which contained the information required; 48,054 pages were automatically scraped. Out of all pages, 1571 were found for the Computer Science field, 1041 for Economics, 3004 for Engineering, 265 for Environmental Sciences, and 451 for Mathematics. Data were stored in JSON format and further processing was required to clean it before using it as an input in data analysis instruments.

4. Methodology

Consequently, the input in this investigation is represented by a considerable amount of textual data. To be more precise, a collection of documents representing descriptions of research job positions constitutes the source of information in this study. Usually, in the text mining literature, a collection of documents is labelled corpus. In order to extract information from it, as we know from big data theory, an interdisciplinary approach is recommended. Therefore, tools of informatics, programming, statistics, data analysis as well as the domain experts to evaluate and validate the outputs, are required.

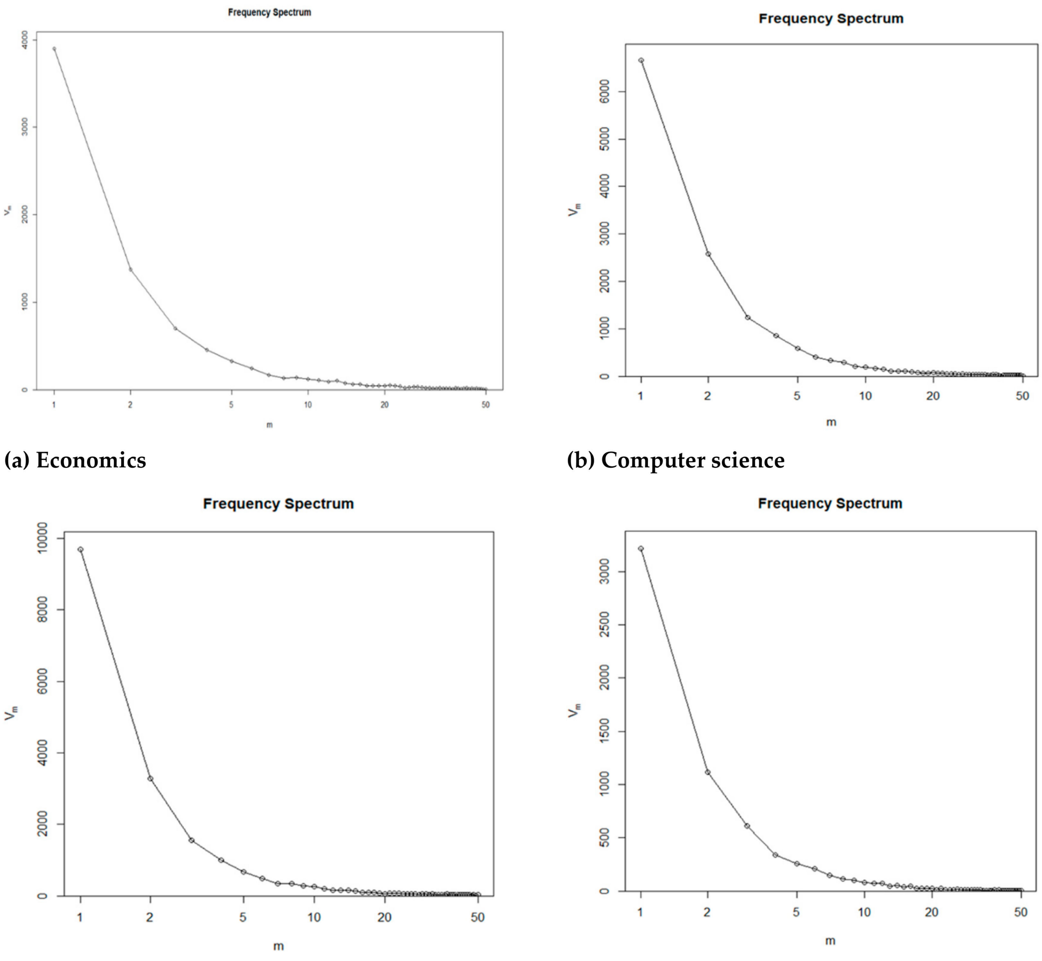
As mentioned in the previous section, the findings of this study are based on the investigation of five different corpuses. Text mining was performed with tm library in R [54–56] and mainly used for text summarization. Other concepts employed in the text mining literature, besides the corpus, are document, token and lexicon or dictionary. In our investigation, the document is a research job offer; the tokens are the fundamentals units of analysis, represented by individual words. The lexicon or the dictionary consists of all unique words in a specific corpus, meanwhile the corpus size indicates the total number of words used [57]. Large texts are analyzed by computational methods based on statistical concepts. In order to use such methods, data transformation is required. This stage involves building structured representations similar to those employed in classical data mining such as matrices. The final results will consist of a matrix known as a document-term matrix whose elements are numerical, representing word frequency. The rows are the documents and the columns are the tokens [55,58]. At first, before completing data cleaning, this matrix is very large and extremely sparse. Table 1 emphasizes the vocabulary size and the corpus size before undertaking this pre-processing action. We have to mention that these numbers were computed after conducting some preliminary cleaning operations such as: eliminating extra white spaces, removing conjunctions and prepositions (stop words), removing punctuation, and converting to lower case.

There are universal regularities characterizing word frequency distribution, of which the best known is the theory of the minimum effort (Zipfs' law) [59,60]. This theory is used in the literature to compare and asses the quality of the informational content of a text. Zipfs' law states that the frequency of occurrence of a word is approximately an inverse power law function of its rank. The parameters appearing in this law will characterize the diversity of the vocabulary. In order to understand if the five corpuses investigated in this paper depict this universal law and moreover to see if there are significant differences between them in terms of vocabulary richness, we summarized the frequency distribution through a frequency spectrum. This involves computing Vm, representing the number of words occurring m times in the corpus. We have plotted the frequency spectrum for the first 50 elements [60,61]. The main conclusion we can draw from this representation is that the corpuses analyzed in this paper follows a typical frequency pattern suggested by the plot in Figure 1.

Table 1. Dimension of the corpus.

Field	Before Data Cleaning		After Data Cleaning	
	Vocabulary Size	Corpus Size	Vocabulary Size	Corpus Size
Engineering	21570	363158	608	108634
Economics	9280	126301	625	45295
Computer Science	16306	265725	851	84362
Environmental sciences	6939	41322	797	14199
Mathematics	7199	53841	587	17260

Source: authors' computation.



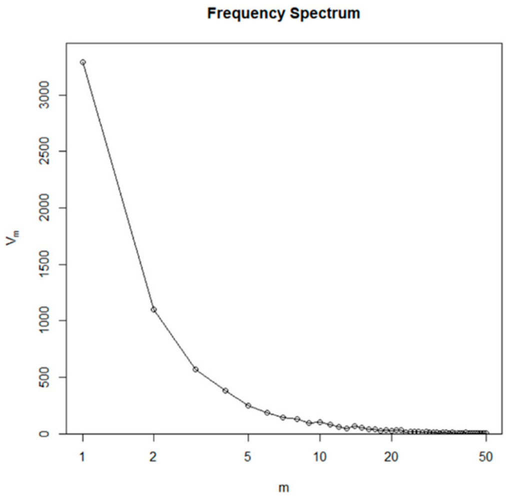
(a) Economics

(b) Computer science

(c) Engineering

(d) Environmental studies

Figure 1. Cont.



(e) Mathematics

Figure 1. Frequency spectrum.

Generally, in text mining, the data transformation process which leads to a frequency matrix is followed by data cleaning. Typically, this stage involves pre-processing the corpus through the following operations: eliminating extra white spaces, removing conjunctions and prepositions (stop words), removing punctuation, converting to lower case, and application of a stemming algorithm. As we mentioned before, we made use of all except the last one. We decided not to use a stemming algorithm, which removes word' suffixes with the purpose of dimension reduction. As we exemplified in the next section, retrieving the radicals of some words can lead to the loss of relevant information. We also defined and eliminated non-relevant words such as: applicant, required, email, and position, which are common in the textual data coming from job advertisements, with no informational value in the context of our investigation.

In order to reduce the size of the document-term matrix, we used two different procedures. The first one excludes the sparse terms and the second one eliminates the words appearing in almost all the documents.

The maximal allowed sparsity was set to 0.98. This means that those columns associated to very infrequent words were dropped. The sparsity was computed for each term by the formula:

$$sparsity_i = 1 - \frac{n_i}{N} \quad (1)$$

where n_i represents the number of occurrences for term i and N is the total number of documents in the corpus. In our analysis were kept only those words with a sparse factor of less than the threshold of 0.98. We consider that rare terms do not contribute to our findings given that we are interested in finding which are the most required skills.

On the opposite side, the document-term matrix contains words occurring in almost all the documents. Such words are not necessarily related to our topic, they are rather common words specific to the job posts. In this case, the elimination is made switching from a weighting system based on term frequency to a scheme known as inverse document frequency emphasizing the words with higher discriminative power [57,58,62].

The inverse document frequency for a specific term is computed by the formula:

$$idf_i = log_2\left(\frac{N}{d_i}\right) \quad (2)$$

where N represents the size of the corpus and d_i represents the number of documents where the term i appears [62].

Generally, in text mining, the statistical measure used to evaluate the importance of a certain term is given by the tf-idf which stands for term frequency-inverse document frequency. The importance increases proportionally to the number of occurrences in the document but is counterbalanced by the frequency of the word in the entire collection of documents. This implies normalization of a term frequency using the document length measured by the total number of words in the document (tf_i). Therefore, the tf-idf is computed by [57,58]:

$$tf\_idf_i = tf_i * idf_i \tag{3}$$

We have computed this statistic for each term and we discarded all the terms obtaining a value smaller than the first quantile. Following these procedures, we have significantly reduced the size of the term document matrix. The dimensions of each corpus and the summary statistics for the tf-idf values are given in Table 2.

Table 2. Term frequency—descriptive statistics.

Corpus	Length (No. of Documents)	Tf-Idf Statistics			
		1st Qu.	Median	Mean	3rd Qu.
Engineering	3004	0.052	0.066	0.069	0.082
Economics	1043	0.043	0.059	0.0616	0.074
Computer science	1571	0.034	0.042	0.044	0.052
Environmental sciences	265	0.033	0.043	0.05	0.06
Mathematics	451	0.045	0.056	0.064	0.074

Source: authors' computation.

The implications of these operations on the dimension of each corpus are revealed in Table 2 as well as in Appendix A. As illustrated in Table 1, the vocabulary size sharply decreased, and also the variance of this measure across the research fields significantly declined. Among all research fields, computer science depicts the largest vocabulary. As we will find in the next section, this could be explained by the diversity of the domains where computational methods are required. The representation included in Appendix A plots the top ten most frequent terms before and after undertaking the text cleaning, revealing that the final matrices do not include terms without informational value.

These collections of documents were used to identify the skills and knowledge required in the research sector for five different fields.

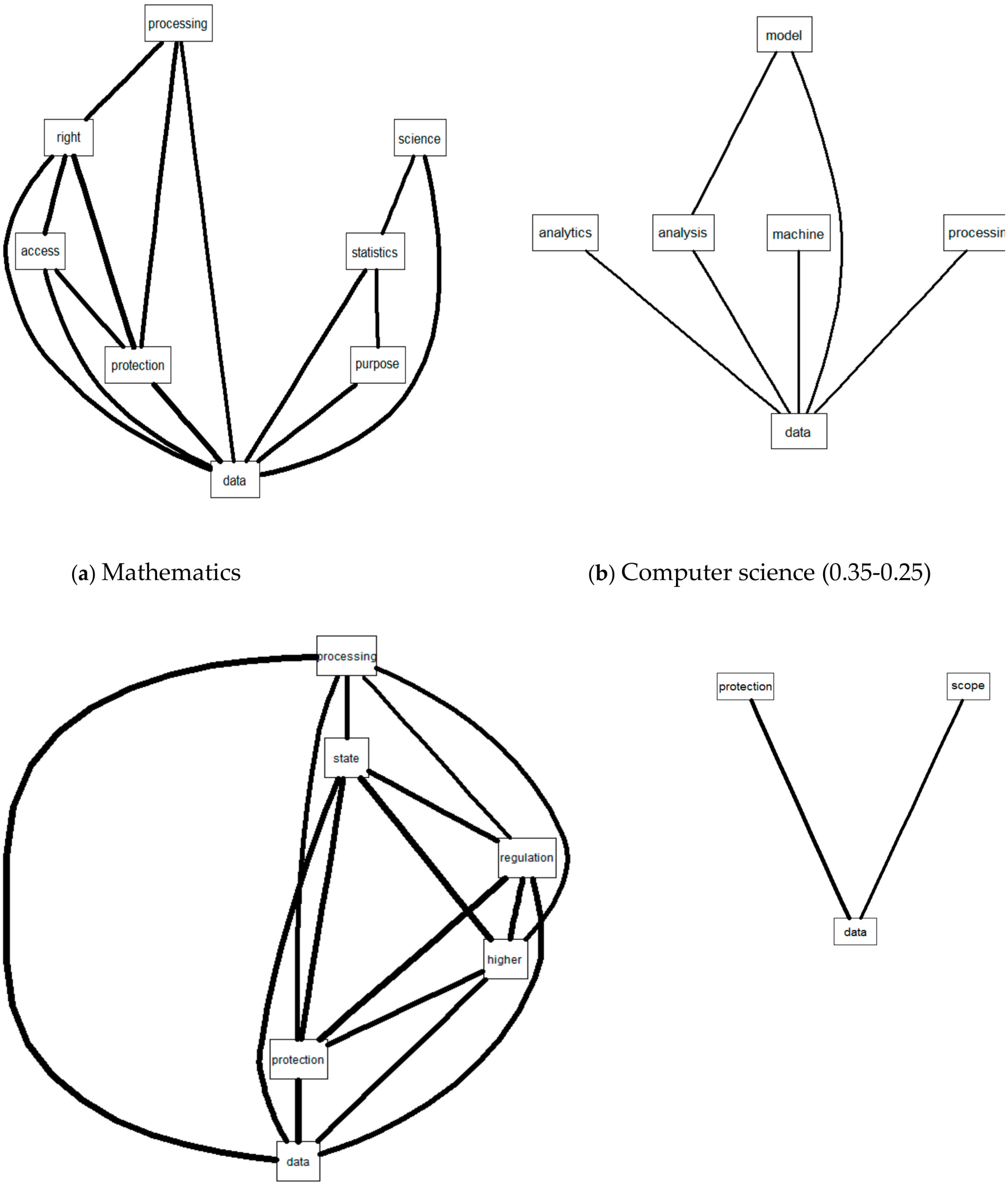
In the next section, the findings are represented via word clouds, a visual instrument highlighting the most frequently used terms in the advertisements of the vacancies or in the calls for applications. The frequency of a certain word is computed by the sum of the column it represents in the document-term matrices obtained after the cleaning and transformation process. The word clouds we inserted in the paper use the top 100 most frequent words [63,64].

For a better understanding of the word clouds, we have also extracted and represented the associations encountered for different terms. In essence, the correlations among those words are computed indicating the share of co-occurrences. This tool allows us to draw the context in which those terms are used.

5. Results

The main findings are extracted from the word cloud representation associated to each research field. The most obvious conclusion that can be drawn at first glance from the word frequency visualization is related to the interdisciplinary dimension of the research activity. This facet may also be a consequence of the fact that many of the job posts include more than one research field. Further research should deal with this issue using classification methods.

Without exception, the aspects related to “data” are very often specified in job descriptions and/or requirements. In order to understand the context, we analyzed the associations of this word and we found that it co-occurs with ”protection”, “statistics”, “science”, and “processing”. In the Figure 2 inserted below are represented only the correlations exceeding a threshold of 0.5. For computer science for example, the highest correlation is of 0.35. In this case, it co-occurs with terms such as “analytics”, “analysis”, “processing”, “machine” or “model”. We can conclude that at least in mathematics, engineering and computer science, data mining or data processing skills are frequently required.



(**a**) Mathematics

(**b**) Computer science (0.35-0.25)

(**c**) Engineering

(**d**) Economics

Figure 2. Visualization of the correlations for term “data”. Source: authors’ computation.

Besides the term “data”, another term common to all fields, appearing with a high frequency, is “model”. In computer science, the highest correlations are found for “simulation” (0.39) and “scientist”, pointing again towards data science skills. For economics, the most significant association was found with “energy” (0.39) and “analyses” (0.37). For the documents extracted from the engineering field, the first association (0.34) was found for “numerical”, showing that the advertisements of the vacancies including the term “model” also contain “numerical”. The corpus obtained for environmental sciences revealed many terms co-occurring with the term “model”. We consider that the relevant ones could be “simulation” (0.67), “dynamics” (0.67) or “surveys” (0.64). For the mathematics field, the words that could explain the context in which it occurs are “simulation” (0.29) and “optimization” (0.24). Hence, the term “model” indicates requirements related to analytical skills. This conclusion is also supported by the frequency of the term “analysis” which is easily detectable in all five figures.

Additionally, the word “university” plays an important role in the documents we have analyzed, and this is due to the fact that most of the posts are coming from universities and implicitly the name of the university is mentioned in the job description. Regularly, a position in a university also implies teaching activity, which is reflected by our word cloud through terms such as “courses”, “teaching”, “assistant”, and “professor”.

The term “management” is common to all five-word clouds but it is difficult to summarize its correlations due to the fact it is related to a wide variety of aspects. For instance, no matter the field, it is associated with the following terms: “time”, “risk”, “financial”, “supply chain”, “organizational”, “project”, ”industrial”, ”quality”, ”resources”, ”financial”, ”team”, and ”strategic”. So, data processing and handling, teaching and management skills could be considered core competences for R&D professionals that transcend all the analyzed fields.

Specific skills required within the five analyzed fields are presented in the following paragraphs. For the vacancies associated with engineering field (Figure 3), words such as ”physics”, ”energy”, ”materials”, ”mechanical”, ”electrical”, ”mathematics”, and ”electronics” are keywords for the technical knowledge required. On the other hand, the presence of the words ”communication” or ”language” unveil a different facet of the research activity which requires good oral and written communication skills. The term ”language” is mostly associated with ”English” (0.43) and ”foreign” (0.4), emphasizing that ”English” is used as a scientific and research language. The research topics within the project calls could be very heterogeneous and it is indicated by terms such as “medical”, “sustainability” or “environmental”. Another dimension that could be extracted from the representation is about IT-related competences. For instance, the term “software” acquired a significant frequency being correlated with “programming”, “testing” or “computer”.

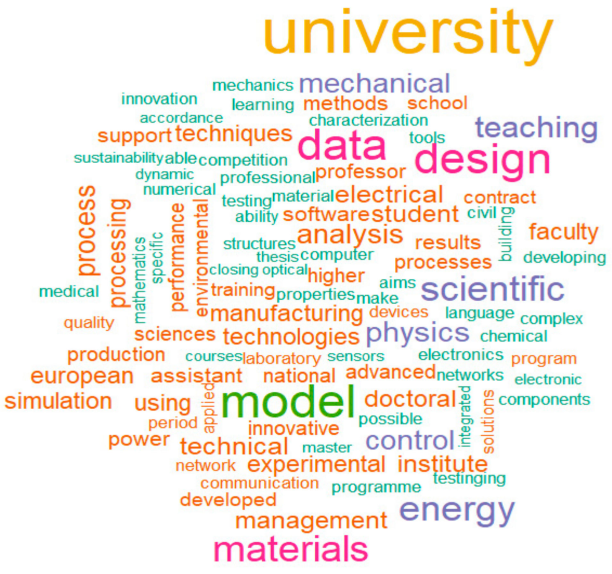


Figure 3. Engineering. Source: authors’ computation.

The representation built for the environmental sciences field (Figure 4) highlights more keywords related to the research topics of the projects being advertised: “water”, ”climate”, ”ocean”, ”earth”,

″biology″, ″change″, ″ecosystem″, and ″sustainability″. Therefore, this investigation also depicts the topical research areas.

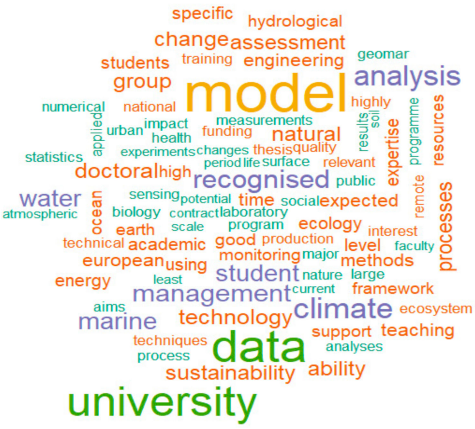


Figure 4. Environmental sciences. Source: authors' computation.

As mentioned in the methodology section, we did not use a stemming algorithm because we consider that relevant information could be lost. For example, in the word cloud associated to Mathematics field (Figure 5), one can distinguish the term "science" but also the plural form "sciences". Employing a stemming algorithm will lead to the elimination of the second one together with the absorption of its frequency by the radical of the word. However, if we analyze the associations for these words, we see that they are indicating different aspects. "Science" is rather related to data science meanwhile the prevalence of the word sciences comes from collocations like natural sciences or social sciences. Among technical skills, related to this research field are very well represented: "probability theory", "differential equations", and "physics". IT skills are now reflected by terms like "computer" or "program". Moreover, an interesting aspect also related to softwareengineering skill is highlighted by the associations found for the word "machine". The word "learning" is frequently associated with it (0.87), depicting that the candidates should be able to implement different machine learning algorithms. Among the top 100 terms, "team" is present, emphasizing that team-work is essential for research activity. The word "language" is associated with diplomas and certificates (0.4), showing that good communication in a foreign language is required.

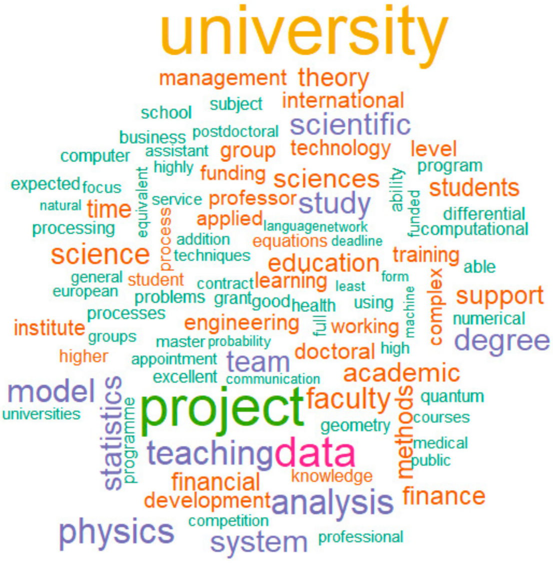


Figure 5. Mathematics. Source: authors' computation.

For the first time, the word cloud representing the Economics field (Figure 6) highlights an important dimension of the research activity, the publication of which is an outcome but also a selection criterion. This is summarized by “article”, “publication” or “journals” words situated among first 100 most frequent. The magnitude of the term “business” is coming from the requirements related to the candidate studies: Master’s degree or PhD or equivalent relevant experience in the field of finance or management/business administration are required. This is why the two words most correlated with “business” are “school” and “administration”.

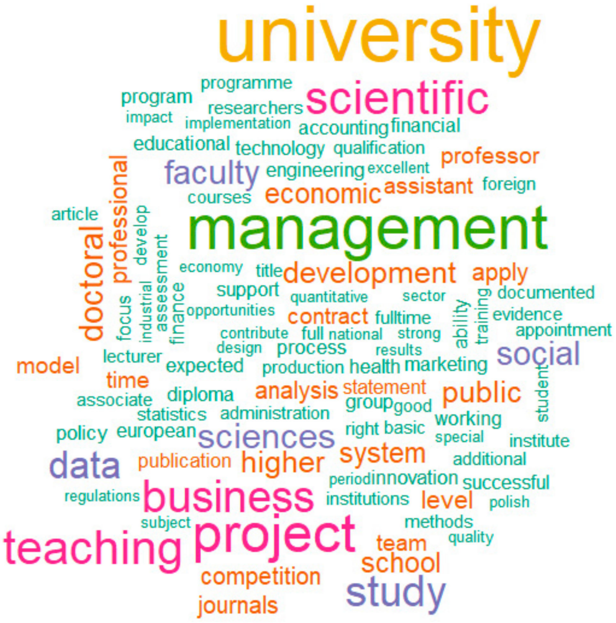


Figure 6. Economics. Source: authors’ computation.

For the computer science field (Figure 7), terms such as “deep”, “machine”, “algorithms”, “learning”, “digital”, “computing”, and “intelligence”, are keywords that could be anticipated. Therefore, successful candidates should have technical skills related to machine learning or its new area known as deep learning which is mostly based on artificial neural networks.

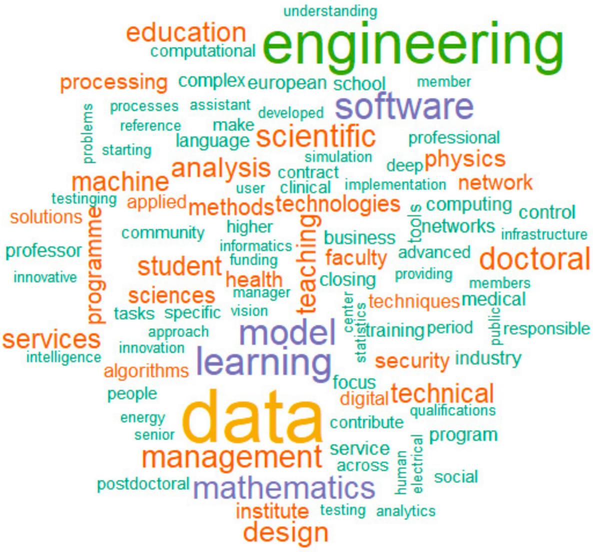


Figure 7. Computer science. Source: authors’ computation.

The appearance of the words "medical", "health", "human", "social", and "clinical" could indicate topical research themes for the period we are analyzing. This is also due to the that fact that machine learning algorithms are often applied to medical data sets. The term innovative is pointing towards technologies and solutions that should be developed in the projects for which the research positions are opened.

6. Conclusions

Our article proposes a text mining approach in order to identify the mix of skills that are required from first stage researchers in Europe. The analysis was applied on job vacancies advertisements extracted from the EURAXESS platform for selected research fields: Computer Science, Economics, Engineering, Environmental Science and Mathematics.

First, the results of our analysis can be utilized by educational institutions that contribute to skill formation of future R&D professionals. Second, the results are relevant for early career researchers, PhD candidates and career guidance providers who can better understand the entry level research labour market in terms of skills and demands at the workplace. Third, the results can be useful for R&D companies themselves which can benefit from the overview of the main developments characterizing various research fields. They can better develop effective human resources policies in order to attract, develop and retain the right mix of skills.

Another important conclusion is that text mining analysis of job vacancies advertisements is very useful for identifying the mix of skills required by employers in R&D sector from first stage researchers. Our results show that data handling and processing skills are essential for early career researchers, irrespective of their research field. Also, first stage research positions are connected to universities and include teaching activities to a great extent. Management of time, risks, projects and resources plays an important part in the job requirements included in the analyzed advertisements. Considering the obtained word clouds, we can conclude that R&D professionals face rapidly changing demands determined by new technologies developments and environmental challenges. IT skills have also been highlighted by the word clouds in all research fields. In fact, one could see that nowadays first stage research positions include aspects that are embedded in all types of R&D career orientation (technical orientation, manager orientation, project orientation, technical transfer orientation and entrepreneur orientation). It indicates a diversification of job tasks for early career researchers in line with the increased interdisciplinary and transformations of the research sector.

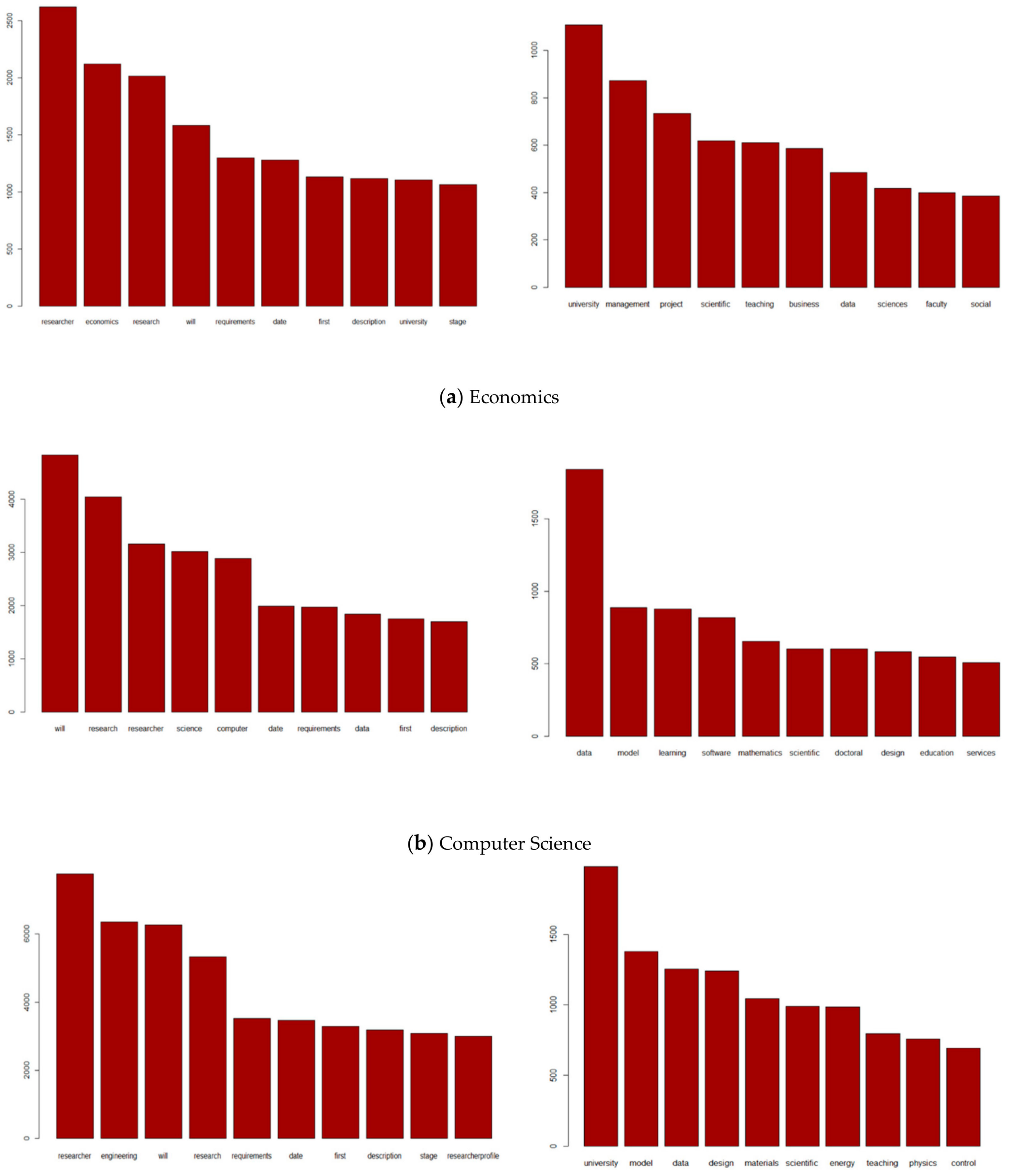
Moreover, the proposed methodological approach is very helpful for exploring specific, technical skills which are much more complicated to be assessed and which are usually studied via in-depth sector level analysis. The main limitations of the study are related to the short time span covered by the gathered job advertisements, limited number of research fields that have been analyzed and the fact that the EURAXESS platform is used mostly by universities and less by research companies. In our future research, we plan to study the dynamics of the skill needs for R&D professionals, to compare entry level positions with more advanced ones, to perform cross-country comparisons, and to include more research fields in our analysis. Further research will extend the collection of textual data to other different research fields in order to extract specific latent variables known as topics. These represent a cluster of words with similar meanings and could lead to a classification of our documents according to the prevalence of topics that describe each document.

Author Contributions: Conceptualization, data processing and manuscript writing M.M.M.M.; manuscript writing and editing C.M and A.M.Z.; data gathering and manuscript writing T.M.G.

Funding: This paper has been elaborated in the NUCLEU Program 19N/2019, funded by Romanian Research and Innovation Ministry, projects PN 19130301 and PN 19130302. The APC was funded by PN 19130301.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

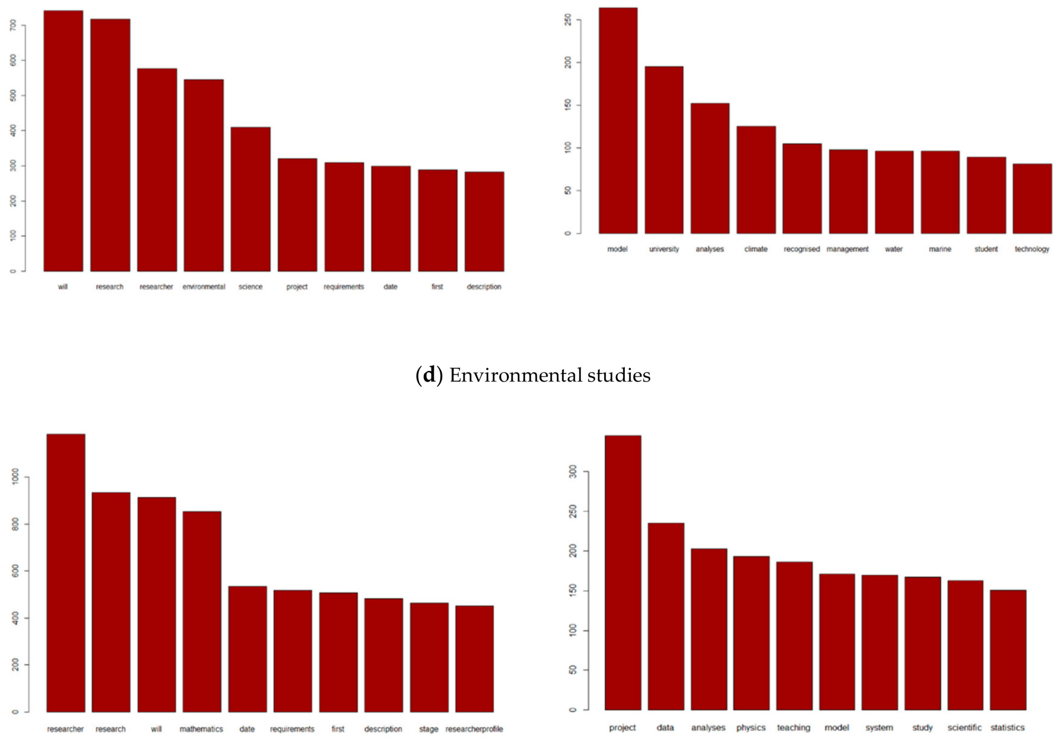


(**a**) Economics

(**b**) Computer Science

(**c**) Engineering

Figure A1. *Cont.*



(**d**) Environmental studies

(**e**) Mathematics

Figure A1. Top ten most frequent words before and after data cleaning. Source: authors' computation.

References

1. Vuong, Q.H. The (ir)rational consideration of the cost of science in transition economies. *Nat. Hum. Behav.* **2018**, *2*, 5. [CrossRef]
2. European Commission. *European Research Area Progress Report 2018*; European Commission: Brussels, Belgium, 2019.
3. Roberts, K. The entry into employment: An approach towards a general theory. *Sociol. Rev.* **1968**, *16*, 165–184. [CrossRef]
4. Aryee, S. Career orientations, perceptions of rewarded activity, and career strategies among R&D professionals. *JET-M* **1992**, *9*, 61–82.
5. Allen, T.J.; Katz, R. The dual ladder: Motivational solution or managerial delusion? *R&D Manag.* **1986**, *16*, 185–197.
6. Dalton, G.W.; Thomson, P.H. *Novations: Strategies for Career Management*; Scott Foresman: Glenview, IL, USA, 1986.
7. Chen, T.Y.; Chang, P.L.; Yeh, C.W. Square of correspondence between career needs and career development programs for R&D personnel. *J. High Technol. Manag. Res.* **2003**, *14*, 189–211.
8. Chen, T.Y.; Chang, P.L.; Yeh, C.W. A study of career needs, career development programs, job satisfaction and the turnover intentions of R&D personnel. *Career Dev. Int.* **2004**, *9*, 424–437.
9. Debackere, K.; Buyens, D.; Vandenbossche, T. Strategic career development for R&D professionals: Lessons from field research. *Technovation* **1997**, *17*, 53–62.
10. Burack, E.H.; Burack, M.D.; Miller, D.M.; Morgan, K. New paradigm approaches in strategic human resource management. *Group Organ. Stud.* **1994**, *19*, 141–159. [CrossRef]
11. Wheelwright, S.C.; Clark, K.B. *Revolutionizing Product Development*; The Free Press: New York, NY, USA, 1992.
12. García-Sánchez, E.; Guerrero-Villegas, J.; Aguilera-Caracuel, J. How Do Technological Skills Improve Reverse Logistics? The Moderating Role of Top Management Support in Information Technology Use and Innovativeness. *Sustainability* **2019**, *11*, 58. [CrossRef]
13. Igbaria, M.; Kassicieh, S.K.; Silver, M. Career orientations and career success among research, and development and engineering professionals. *J. Eng. Technol. Manag.* **1999**, *16*, 29–54. [CrossRef]

14. Garden, A.M. Career orientations of software developers in a sample of high tech companies. *R&D Manag.* **1990**, *20*, 337–353.
15. OECD. *Workforce Skills and Innovation: An Overview of Major Themes in the Literature*; OECD: Paris, France, 2011.
16. National Science Board. *Research & Development, Innovation, and the Science and Engineering Workforce: A Companion to Science and Engineering Indicators 2012*; NSB-12-03; National Science Foundation: Arlington, VA, USA, 2012.
17. Davidescu, A.A.; Paul, A.M.V.; Gogonea, R.-M.; Zaharia, M. Evaluating Romanian Eco-Innovation Performances in European Context. *Sustainability* **2015**, *7*, 12723–12757. [CrossRef]
18. Gerport, T.J.; Domsch, M.; Keller, R.T. Career orientations on different countries and companies: An empirical investigation of West Germany, British and US industrial R&D professionals. *J. Manag. Stud.* **1988**, *25*, 439–462.
19. Aryee, S.; Leong, C.C. Career orientations and work outcomes among industrial R&D professionals. *Group Organ. Stud.* **1991**, *16*, 193–205.
20. Kim, Y.; Cha, J. Career orientations of R&D professionals in Korea. *R&D Manag.* **2000**, *30*(2), 121–137.
21. Igbaria, M.; Greenhaus, J.H.; Parasuraman, S. Career orientation of MIS employees: An empirical analysis. *MIS Q.* **1991**, *15*, 151–169. [CrossRef]
22. Schein, E.H. How career anchors hold executives to their career paths. *Personnel* **1975**, *52*, 11–24.
23. Turpin, T.; Deville, A. Occupational roles and expectations of research scientists and research managers in scientific research institutions. *R&D Manag.* **1995**, *25*, 141–157.
24. McCormick, K. Career paths, technological obsolescence and skills formation: R&D staff in Britain and Japan. *R&D Manag.* **1995**, *25*, 197–211.
25. Cable, D.M.; Judge, T.A. Person-organization fit, job choice decisions and organizational entry. *Organ. Behav. Hum. Decis. Process* **1996**, *67*, 294–311. [CrossRef]
26. Kristof, A. Person-organization fit: An integrative review of its conceptualizations, measurement, and implication. *PPSych* **1996**, *49*, 1–49. [CrossRef]
27. Yi, J.C.; Kang-Yi, C.D.; Burton, F.; Chen, H.D. Predictive Analytics Approach to Improve and Sustain College Students' Non-Cognitive Skills and Their Educational Outcome. *Sustainability* **2018**, *10*, 4012. [CrossRef]
28. Zhang, L.; Guo, X.; Lei, Z.; Lim, M.K. Social Network Analysis of Sustainable Human Resource Management from the Employee Training's Perspective. *Sustainability* **2019**, *11*, 380. [CrossRef]
29. OECD. *Skills for Innovation and Research*; OECD Publishing: Paris, France, 2011.
30. Auriol, L. *Careers of Doctorate Holders: Employment and Mobility Patterns*; OECD STI Working Paper 2010/4; OECD: Paris, France, 2010; 8p.
31. Forfas. *The Role of PhDs in the Smart Economy*; Advisory Council for Science, Technology and Innovation: Dublin, Ireland, 2009; Available online: http://www.sciencecouncil.ie/media/asc091215_role_of_phds.pdf (accessed on 3 April 2019).
32. Mietzner, D.; Kamprath, M. A Competence Portfolio for Professionals in Creative Industries. *Creat. Innov. Manag.* **2013**, *22*, 280–294. [CrossRef]
33. O*Net OnLine. Available online: https://www.onetonline.org/ (accessed on 30 March 2019).
34. Skills Panorama. Available online: https://skillspanorama.cedefop.europa.eu/en/glossary/s (accessed on 20 March 2019).
35. Sasso, S.; Ritzen, J. Sectoral cognitive skills, R&D, and productivity: A cross-country cross-sector analysis. *Educ. Econ.* **2019**, *27*, 35–51.
36. Piva, M.; Vivarelli, M. The role of skills as a major driver of corporate R&D. *Int. J. Manpow.* **2009**, *30*, 835–852.
37. Mariani, M. What determines technological hits? Geography versus firms competencies. *Res. Policy* **2004**, *33*, 1565–1582. [CrossRef]
38. Michelacci, C. Low returns in R&D due to the lack of entrepreneurial skills. *Econ. J.* **2003**, *113*, 207–225.
39. Quelin, B. Core Competencies, R&D Management and Partnerships. *Eur. Manag. J.* **2000**, *18*, 476–487.
40. CEDEFOP. *Future Skill Needs for the Green Economy*; Publications Office of the European Union: Luxembourg, Belgium, 2009.
41. European Commission. *Providing Researchers with the Skills and Competencies They Need to Practice Open Science*, Open Science Skills Working Group Report; European Commission: Brussels, Belgium, 2017.
42. Gardiner, A.; Aasheim, C.; Rutner, P.; Williams, S. Skill Requirements in Big Data: A Content Analysis of Job Advertisements. *J. Comput. Inf. Syst.* **2018**, *58*, 374–384. [CrossRef]
43. Wowczko, I.A. Skills and vacancy analysis with data mining techniques. *Informatics* **2015**, *2*, 31–49. [CrossRef]

44. Kennan, M.A.; Willard, P.; Cecez-Kecmanovic, D.; Wilson, C.S. IS knowledge and skills sought by employers: A content analysis of Australian IS early career online job advertisements. *Aust. J. Inf. Syst.* **2008**, *15*, 1168044.
45. Yang, Q.; Zhang, X.; Du, X.; Bielefield, A.; Liu, Y. Current market demand for core competencies of librarianship—A text mining study of American Library Association's advertisements from 2009 through 2014. *Appl. Sci.* **2016**, *6*, 48. [CrossRef]
46. Kobayashi, V.B.; Mol, S.T.; Berkers, H.A.; Kismihok, G.; Den Hartog, D.N. Text mining in organizational research. *Organ. Res. Methods* **2018**, *21*, 733–765. [CrossRef] [PubMed]
47. Karakatsanis, I.; AlKhader, W.; MacCrory, F.; Alibasic, A.; Omar, M.A.; Aung, Z.; Woon, W.L. Data mining approach to monitoring the requirements of the job market: A case study. *Inf. Syst.* **2017**, *65*, 1–6. [CrossRef]
48. Vermna, A.; Yurov, K.M.; Lane, P.L.; Yurova, Y.V. An investigation of skill requirements for business and data analytics positions: A content analysis of job advertisements. *J. Educ. Bus.* **2019**, *94*, 243–250. [CrossRef]
49. European Commission. Euraxess. 2019. Available online: https://euraxess.ec.europa.eu/ (accessed on 15 May 2019).
50. Scrapinghub Ltd. Scrapy. 2019. Available online: https://docs.scrapy.org/en/latest/intro/overview.html (accessed on 25 April 2018).
51. Zhang, S.; Feick, R. Understanding public opinions from geosocial media. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 74. [CrossRef]
52. Hernandez-Suarez, A.; Sanchez-Perez, G.; Toscano-Medina, K.; Martinez-Hernandez, V.; Perez-Meana, H.; Olivares-Mercado, J.; Sanchez, V. Social sentiment sensor in Twitter for predicting cyber-attacks using $\ell 1$ regularization. *Sensors* **2018**, *18*, 1380. [CrossRef] [PubMed]
53. Boja, C.E.; Herţeliu, C.; Dârdală, M.; Ileanu, B.V. Day of the week submission effect for accepted papers in Physica A, PLOS ONE, Nature and Cell. *Scientometrics* **2018**, *117*, 887–918. [CrossRef]
54. Feinerer, I. Introduction to the tm Package Text Mining in R. 2013. Available online: http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf (accessed on 10 October 2018).
55. Meyer, D.; Hornik, K.; Feinerer, I. Text mining infrastructure in R. *J. Stat. Softw.* **2008**, *25*, 1–54.
56. TM Library R. Available online: https://cran.r-project.org/web/packages/tm/tm.pdf (accessed on 20 February 2019).
57. Solka, J.L. Text data mining: Theory and methods. *Stat. Surv.* **2008**, *2*, 94–112. [CrossRef]
58. Srivastava, A.N.; Sahami, M. *Text Mining: Classification, Clustering, and Applications*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2009.
59. Ausloos, M.; Nedic, O.; Fronczak, A.; Fronczak, P. Quantifying the quality of peer reviewers through Zipf's law. *Scientometrics* **2016**, *106*, 347–368. [CrossRef]
60. Baayen, R.H. *Word Frequency Distributions*; Springer Science & Business Media: Berlin, Germany, 2002; Volume 18.
61. Baroni, M.; Evert, S. The zipfR Package for Lexical Statistics: A Tutorial Introduction. 2006. Available online: http://zipfr.r-forge.r-project.org (accessed on 5 June 2019).
62. Hornik, K.; Grün, B. Topicmodels: An R package for fitting topic models. *J. Stat. Softw.* **2011**, *40*, 1–30.
63. Hornik, K.; Meyer, D.; Buchta, C. Slam: Sparse Lightweight Arrays and Matrices. R Package Version 0.1-40. 2016. Available online: https://CRAN.R-project.org/package=slam (accessed on 8 February 2019).
64. Fellows, I. Wordcloud: Word Clouds. R Package Version 2.5. 2014. Available online: http://CRAN.R-project.org/package=wordcloud (accessed on 17 January 2019).