SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2020
ORGANIZATION AND MANAGEMENT SERIES NO. 149

IDENTIFICATION OF DESIRED PROJECT MANAGER COMPETENCE USING TEXT MINING ANALYSIS

Marcin WYSKWARSKI

Instytut Ekonomii i Informatyki, Wydział Organizacji i Zarządzania, Politechnika Śląska;
marcin.wyskwarski@polsl.pl, ORCID: 0000-0003-2004-330X

Purpose: An attempt to identify the competencies of the project manager desired by the employers and to determine whether changes have occurred over time.
Design/methodology/approach: Job offers were automatically downloaded from website with job offers. An analysis of text mining of fragments of offers describing the competence was carried out. The analysis of text mining included initial text processing, creation of corpora of analyzed documents, creation of a document-term matrix, topic modeling algorithm and the use of classic methods derived from data mining.
Findings: The most frequently used words/n-grams and the correlation of selected words/ n-grams with other words/n-grams were presented in the form of drawings. Based on the frequency of words/n-grams and the correlation value, efforts were made to identify the project manager competencies. The topic modeling algorithm was used to generate topics that can also be used to identify expected project manager competencies.
Research limitations/implications: Only offers written in Polish, downloaded from one websites with job offers, which had the phrase "kierownik projektu" ("project manager") in their job title, were analyzed. Data was collected from 09 to 11 April 2018 and from 09 to 11 April 2019.
Practical implications: The method applied can be used by organizations preparing for the profession of a project manager, to modify and better adapt curricula to the needs of the labor market.
Originality/value: Studies have shown that text mining of job offers can, to some extent, help determine the desired project manager competence.

Keywords: text mining, competencies, project manager, word cloud, topic modeling.

Category of paper: research paper, case study.modeling.

1. Introduction

The employee's knowledge, skills, and attitude significantly affect his work. The competences established and specified for individual positions are of great importance when employers determine employment plans, conduct recruitment, select personnel, or conduct promotion policy. Information on competencies currently desired by employers can be useful for people interested in taking up a job in this position, as well as for training organizations to modify and better adapt the training program.

The primary purpose of this work was to determine the competencies desired by employers for the position of a project manager, and whether there have been changes in this area over time. To this end, text mining job offers downloaded from the website were analyzed.

2. Essence of text mining

The popularity of text mining solutions is influenced by the continuous development of the Internet and the accompanying increase in the importance of the opinion-forming role of social networking (Berry & Kogan, 2010) and the fact that 80% of enterprise information is stored in text documents (Tan, 1999). Most text mining solutions do not analyze the meaning of words and sentences. They are only trying to detect certain rules and regularities associated with the occurrence of specific strings (words, phrases).

According to Marti Hearst, text mining is a process aimed at extracting previously unknown information from text resources (Hearst, 1999). Text mining is a fairly young and interdisciplinary field. It originates, among others, from areas such as data mining, information retrieval, text categorization, probabilistic modeling (Kao & Poteet, 2007). These are all kinds of methods, concepts, and algorithms used to process text resources prepared in natural languages. They are implemented in the form of computer programs, which allows automation of processing processes (Gładysz, 2012).

Text mining is used, among others, for obtaining information from text documents, identifying documents containing specific content, automatically creating summaries, reference classifications, referenceless classifications (grouping, clustering), identifying connections, visualizations, and generating answers to a question (Lula, 2005).

In the process of analyzing text documents using text mining, the following three stages can be easily simplified: preliminary text processing, construction of the word frequency matrix, application of classic methods in the field of data mining (Gładysz, 2012).

The effect of initial text processing is to create a word list for each processed document, called a bag of words. Each paper is a separate bag of words. During this process, information about the order of words and relationships between them is omitted, it is assumed that the occurrences of words are independent of each other (Gładysz, 2012). All punctuation marks (e.g., periods, commas, semicolons, dash) and numbers are removed. An essential element is also the removal of words that do not provide additional information (including conjunctions, prepositions, etc.). This operation is performed using the so-called stop-words (language-specific). Creating a sack of words also removes words that are very rare (least frequent) and words very often (called most frequent) - the so-called pruning.

A significant step in step 1 is to transform the words into their basic version (e.g., the verb is converted to infinitive and the noun to singular nominative). This is a so-called word root extraction, lemmatization, and tagging. Stemming is choosing from a given word, a part invariant for all grammatical forms, i.e., core (i.e., removing all kinds of prefixes and suffixes). Lemmatization is a morphological analysis that allows finding the primary structure of a given word, i.e., identifying the lexeme. Lemmatization is possible when a dictionary or an extensive set of inflectional rules is available. Tagging is the selection of an appropriate morph syntactic description (Mirończuk, 2012).

In the second stage, a word frequency matrix is created. For this purpose, the so-called Vector Space Model. In this model, the documents and the words they contain are represented in the form of a matrix. Each processed material is represented by a separate vector that represents the number of occurrences of individual words (Gładysz, 2012; Mirończuk, 2012). This form of representation of documents and related terms is referred to as the document-term matrix. The rows of the matrix represent the processed documents, and the columns contained in them the words. There are different ways to express the space-vector representation of text in an array. Among the commonly used methods of coding information in the matrix, there are Boolean (term) representation, term frequency (TF), inverse-document frequency (IDF), mixed TF-IDF, logarithmic, weighted logarithmic and okapi BM25 (Mirończuk, 2012). The collection of documents on the basis of which a matrix of expression-documents is built is in IT linguistics, often called the body of texts.

In the third stage, various activities are carried out to obtain previously unknown information resulting from the documents analyzed. Apply among others grouping documents (so-called clustering), calculating the frequency of words, calculating the correlation between the terms used to help detect new rules and regularities.

3. Project manager's competence – concept, typology

The issue of competences is an important issue considered in both practice and human resource management theory. The literature on the subject contains many different definitions of abilities. A person's competencies create their personality traits, motivation, skills, self-esteem related to functioning in a group, as well as acquired and used knowledge (Gunz, 1983).

It quite extensively defines the competences of T. Oleksyn. According to him, competencies include internal motivation, talents and predispositions, education and knowledge, experience and practical skills, health, fitness and other psychophysical features relevant for work processes, as well as the attitude and behavior expected at the place of employment and formal authorization to act (Oleksyn, 2006).

There are many different classifications of competences in the literature on the subject. Skills can generally be divided into two main groups, i.e., hard competencies related to a specific job position and soft competencies, i.e., personality traits. Hard competencies are defined as professional, technical, professional, substantive, and functional competences. Soft skills are behavioral, social, and interpersonal competences. (Armstrong et al., 2016).

Lyle and Signe Spencer presented an attractive model of competence. They compared competences to "icebergs." According to this model, the visible part of the mountain (top) is knowledge and skills (they can be easily identified, learned, and developed). The invisible part of the iceberg - located underwater - symbolizes such elements as motivation, behavioral patterns, character traits (they are "hidden" so they are more challenging to modify, develop) (Spencer & Spencer, 1993). The competency classification model proposed by R.L. Katz assumes the division into technical, social, and conceptual competences (Katz, 1974). F. Delamare Le Deist and J. Winterton distinguished cognitive (cognitive), functional, social, and meta competence competences (Le Deist Delamare & Winterton, 2005), and G. Filipowicz social, personal, managerial and professional (Grzegorz, 2014).

In the project management literature, you can find information on the desired competencies of the project manager. First of all, he should have professional knowledge in the area in which the project is implemented, and be an expert in the field of methods and techniques of project management (Pawlak, 2006). Project manager's competences are often presented as a set of knowledge, skills, personality traits, and experience (Pawlak, 2006). The capabilities of the project manager can be divided into the following groups (Wachowiak, Gregorczyk, Grucza, & Ogonek, 2004):

- technical – they enable you to understand the essence of the project and complete the task,
- interpersonal – they are used in establishing and maintaining contacts with people,
- conceptual – they contribute to creative problem solving,

- diagnostic and analytical – enable diagnosis of encountered problems,
- political – ensure the practical environmental impact of the project.

According to K. Piwowar-Sulej, the critical competencies of project managers include (Piwowar-Sulej, 2013):

- high professional qualifications, technical knowledge related to the subject and scope of the project, knowledge of management methods (traditional and modern methodologies),
- ability to set goals, ability to organize the work of a project team,
- independence in the assessment of the facts,
- no prejudices about non-standard working methods,
- well-developed social skills (e.g., negotiation, diplomatic skills, tolerance for a different point of view, marketing approach to the client).

The list of project manager's competences presented by A. Musioł-Urbańczyk consists of 46 items. They were divided into four groups, i.e., professional (19 skills), social (9 competencies), personal (14 competencies), and business competencies (4 competencies). The author's research shows that the key competencies are: communication skills, decision-making skills, leadership, ability to motivate team members, team-building skills, project communication management skills, teamwork, negotiation skills, loyalty, project scope management skills, flexibility (Musioł-Urbańczyk, 2010).

Among the professionals in the field of project management, the following four competence models are trendy (Wyrozębski, 2009):

- IPMA Competence Baseline – a project manager's competence model prepared by the International Project Management Association,
- Project Manager Competency Development Framework – a competency model by the American Project Management Institute,
- National Occupational Standards for Project Management – a competency model created by the British organization Engineering Construction Industry Training Board,
- Professional Competency Standards for Project Management – Australian model of project competences, developed by the Australian Institute for Project Management.

4. Data source and text mining analysis process

Job offers downloaded from the website www.pracuj.pl were analyzed. Data was collected from 09 to 11 April 2018 and from 09 to 11 April 2019. The phrase 'project manager' was used to search for offers. Among the proposals returned by the website www.pracuj.pl there were also those in which another name was used for the position, e.g. "kierownik projektów", "project

manager", "projekt manager," "kierownik programu", "koordynator programu". Some of the ads were written in English. Only ads meeting the following two conditions were selected for analysis:

- the advertisement had to refer to the position “kierownik projektu” (e.g., offers for the position “koordynator projektu”, “kierownik program” etc. we’re not analyzed),
- the content of the announcement was written in Polish.

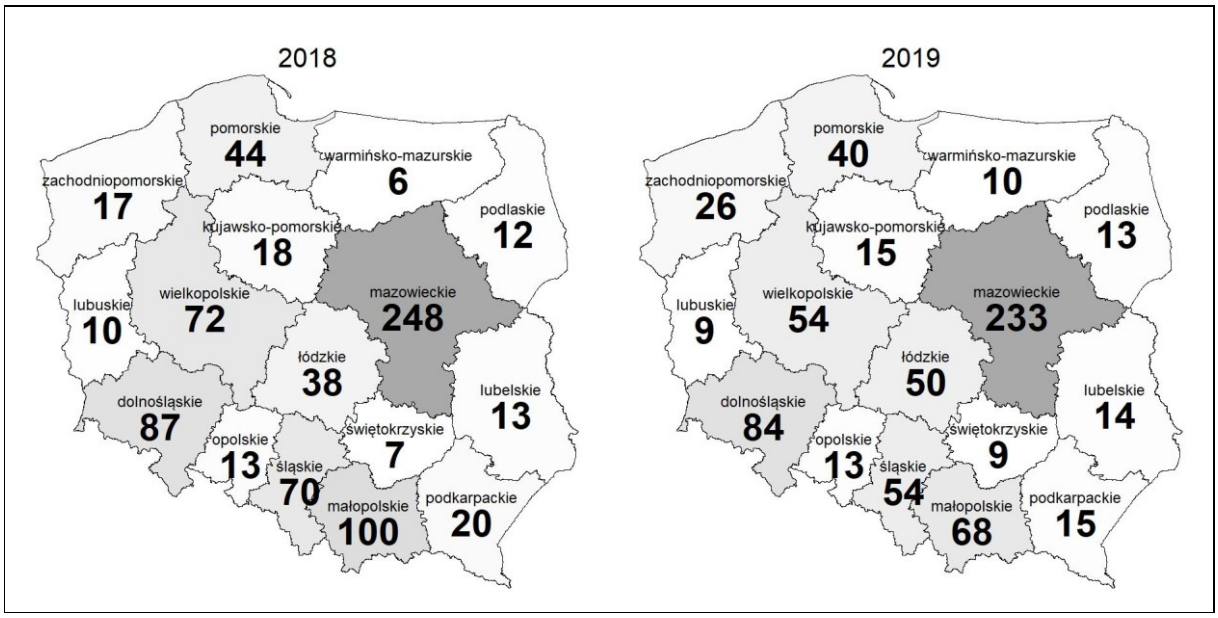

Figure 1. All job offers.

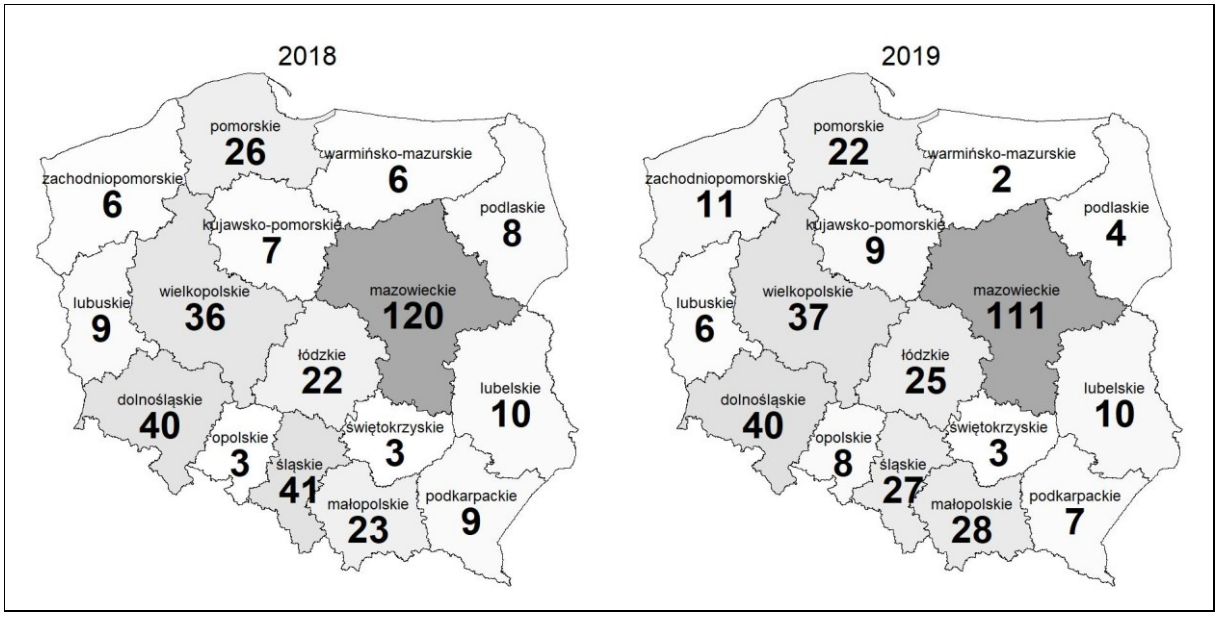

Figure 2. Analysed job offers.

From 775 offers found in April 2018, 369 were selected for analysis. From 707 proposals found in April 2019, 350 were selected for the study. The number of all job offers and those chosen for analysis, broken down by voivodship and year, is presented in Fig. 1. and Fig. 2. A file with the txt extension was created for each offer selected for analysis, in which a fragment

describing the competences of the project manager was saved. The part describing skills was determined using various phrases. The most commonly used terms that have appeared at least ten times are shown in Figure 3 (for 2018 and 2019 together). To enable the analysis of offers for a selected voivodship in a given year, the files were saved in separate folders, broken down by voivodships, and the year the proposal was downloaded (32 folders).

Figure 4 uses the histogram and box chart to distribute the number of words in the created text files for 2018 and 2019 ads. Before counting words, all characters were removed from the files except letters. You can see that these are relatively short documents (the quickest were only eight words, and the average number of words is 50).

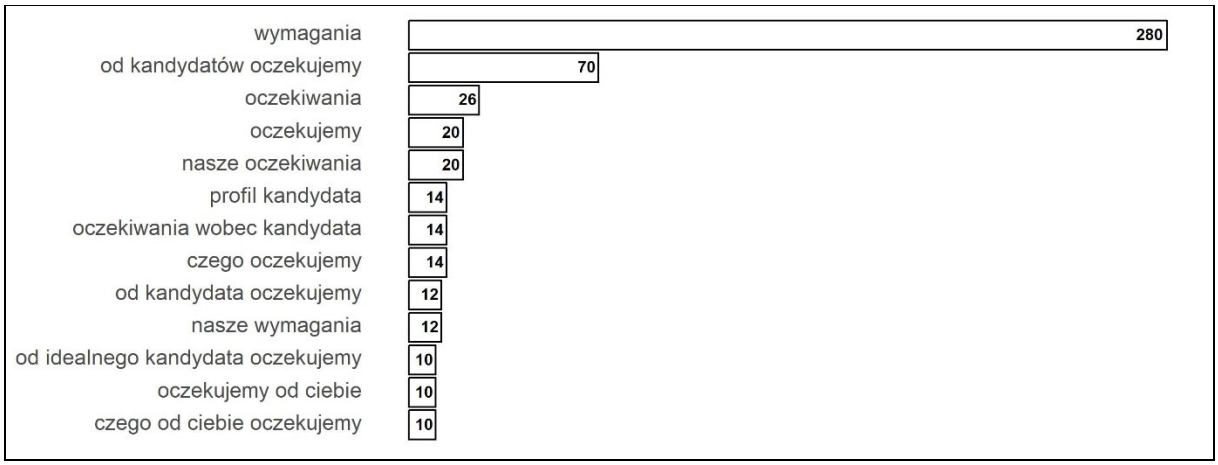

Figure 3. The most commonly used terms for the desired 'competences' of the project manager.

The text mining analysis covered the following three stages:

- preliminary text processing,
- creation of corpora for analyzed documents, construction of the word frequency matrix,
- using classic methods from the data mining area.

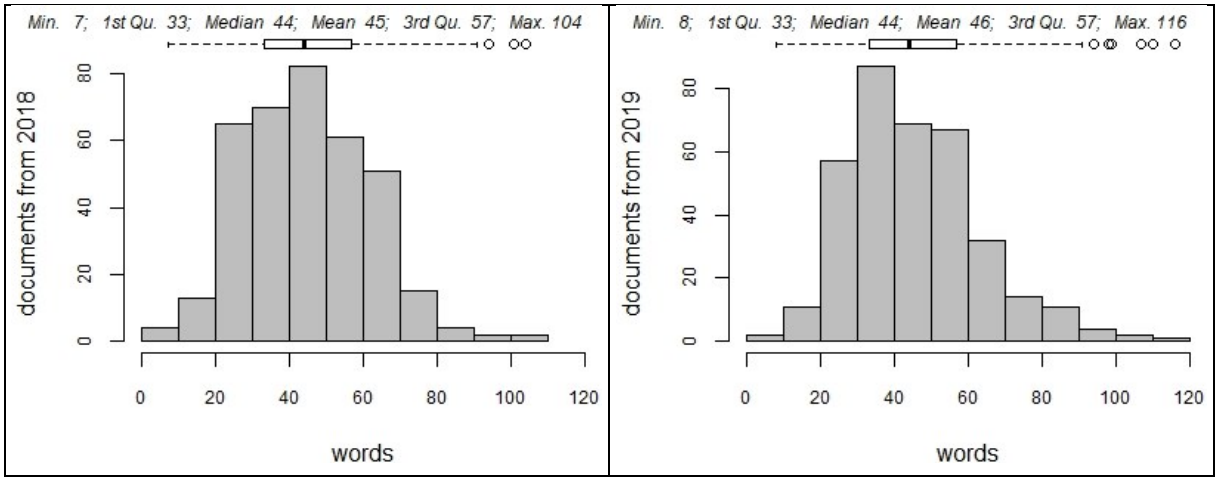

Figure 4. The number of words in created text documents.

As part of the initial processing of each text file converted into so-called sack of words. The Notepad ++ v.7.3.3 and RStudio v.1.0.136 applications were used for this purpose. The demands made it possible to remove all characters except letters, replace uppercase letters with lowercase letters, remove unnecessary words, e.g., conjunctions, prepositions (the so-called stoplist for Polish was used), transform concepts into their basic form, and put words on a separate line.

In order to bring the words back to their basic form, a morphosyntactic dictionary of the Polish language "Polymorphologist 2.1" was used. This dictionary in the form of a text file, containing 4 811 854 lines of text was downloaded from the Github website. After importing the dictionary into the RStudio application, it took the form of a table consisting of columns: primary structure, changed shape, and grammatical markers. Transforming the word to its basic form consisted of searching for it in the "changed form" column and inserting in its place the words from the "basic form" column. If the word was not found in the dictionary, it remained unchanged in the document.

	Terms						
Docs	analityczny	angielski	atut	autocad	biznesowy	branża	budowlany
W_001.txt	0	0	0	0	3	0	0
W_002.txt	0	1	0	0	0	1	0
W_003.txt	0	0	0	0	0	0	0

Figure 5. Extract from the document matrix – expressions for the corpus from 2019.

In the next stage of the analysis, two document bodies were created. The first corpus consisted of offers collected in 2018 and the second one from 2019. Then, for each corpus, a document term matrix was created with a frequency representation of the occurrence of expressions (Term Frequency – TF). When creating the matrix, the so-called Vector Space Model is necessary. Part of the matrix for the corpus including offers collected in 2019, is shown in Figure 5. (the original matrix is 350 x 82, i.e., 350 documents and 82 words).

In the third stage of the analysis, the most common words were found for the corpses formed, which were presented using a bar chart. As part of this stage, the correlation of the selected six words with other words was also calculated. The correlation was calculated using the findAssocs() function based on the standard function cor() available in the R statistical package. The correlation can be from 0 to 1. The value of 1 means that the two words always appear together and in the same amount in documents, and value 0, that the words never appeared together in the analyzed texts.

The last part of the analysis uses the Latent Dirichlet Allocation (LDA) method of popular Topic Modeling algorithm. The technique has been reviewed, among others, by D. Blei, A. Ng, and M. Jordan (Blei, Ng, & Jordan, 2003). The method was used to generate abstract, hidden topics describing the analyzed job offers.

This algorithm assumes that each document is represented by topic division and that each topic is represented as a word breakdown. The identified issues were to facilitate the identification of the desired competencies of the project manager, and enable their possible division into groups (e.g., responsibilities related to team management, obligations arising from conducting construction projects, etc.) and to enable comparison of results for offers from both periods. The work uses the implementation of the LDA algorithm available in the R language package called *topicmodels*.

5. Results of the text mining analysis

The results of the study are presented in graphical form in Figures 6 to 10. Figure 6 shows the fifty most frequently used words (along with the number of their occurrences) in bar graphs in 2018 and 2019 in the form of a bar chart. The words (9 words in each) are marked in gray each list) that appeared only on one of the listings (e.g., the word "obsługa" was in the fifty most common words only in 2018 offers). As you can see, the order of the most used words is similar for the 2018 and 2019 ads. The word "znajomość" appeared first for offers from 2018 and 2019, and the word "experience came 2nd in 2018 and third in 2019. Based on the lists provided, one can try to determine what competencies the employers expect, e.g.,"wykształcenie wyższe"(From the words "wykształcenie" "wysoki"), "znajomość prawa budowlanego" (from the words "prawo" "znajomość"), knowledge of the Office package (from the words:" pakiet", "znajomość", “offic ").

Figure 7. shows the correlation of words with other words appearing in the offers. A relationship of 0 indicates that the two words never occurred together, and a value of 1 that they always occur together (in the same amount) in processed documents. The correlation was presented using point charts divided into 2018 and 2019. Due to the volume, only the six words selected by the author is presented.

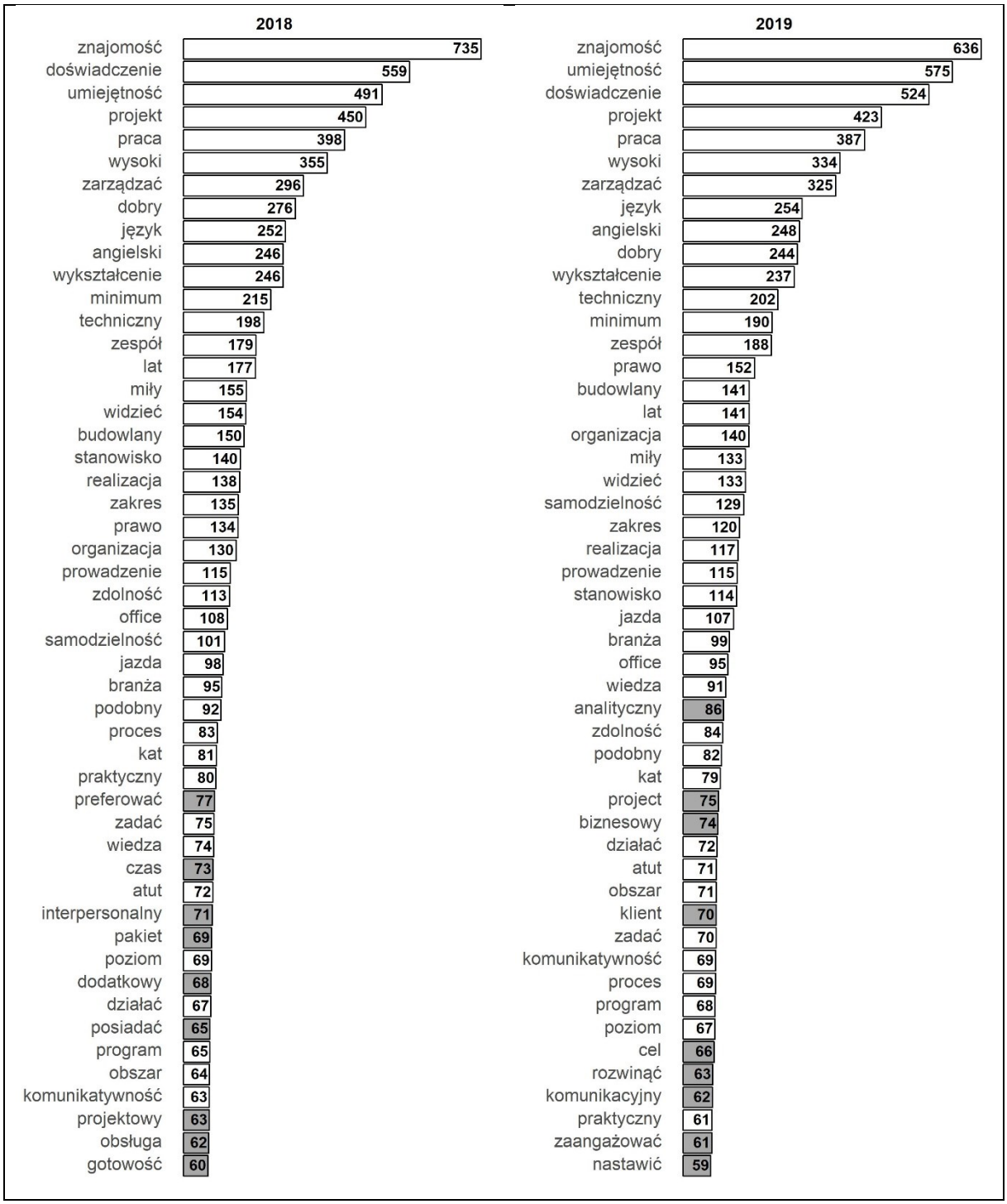

Figure 6. Fifty most commonly used words in ads.

By analyzing charts with correlation values, you can attempt to determine the desired competencies and assess whether changes have occurred over time. For example, from the graph showing the correlation for the word “znajomość”, the following skills could be created: “dobra znajomość języka angielskiego” (from the words “dobry”, “znajomość”, “język”, “angielski”). From the same chart, but only for 2018, it can be determined that "znajomość Office’a będzie dodatkowym atutem" (from the words: "znajomość", "office", "dodatkowy", "atut") and for 2019 that “praktyczna znajomość metodyki Pronce (from the words:

"praktyczny", "znajomość", "metodyka", "prince"). From the chart for the word “umiejętność” for 2018 it can be concluded that „umiejętność pracy pod presją czasu” is valued (from the words „umiejętność”, „praca”, „presja”, „czas”), and for 2019 „umiejętność prowadzenia projektów analitycznych” (from the words: „umiejętność”, „prowadzenie”, „projekt”, „analityczny”).

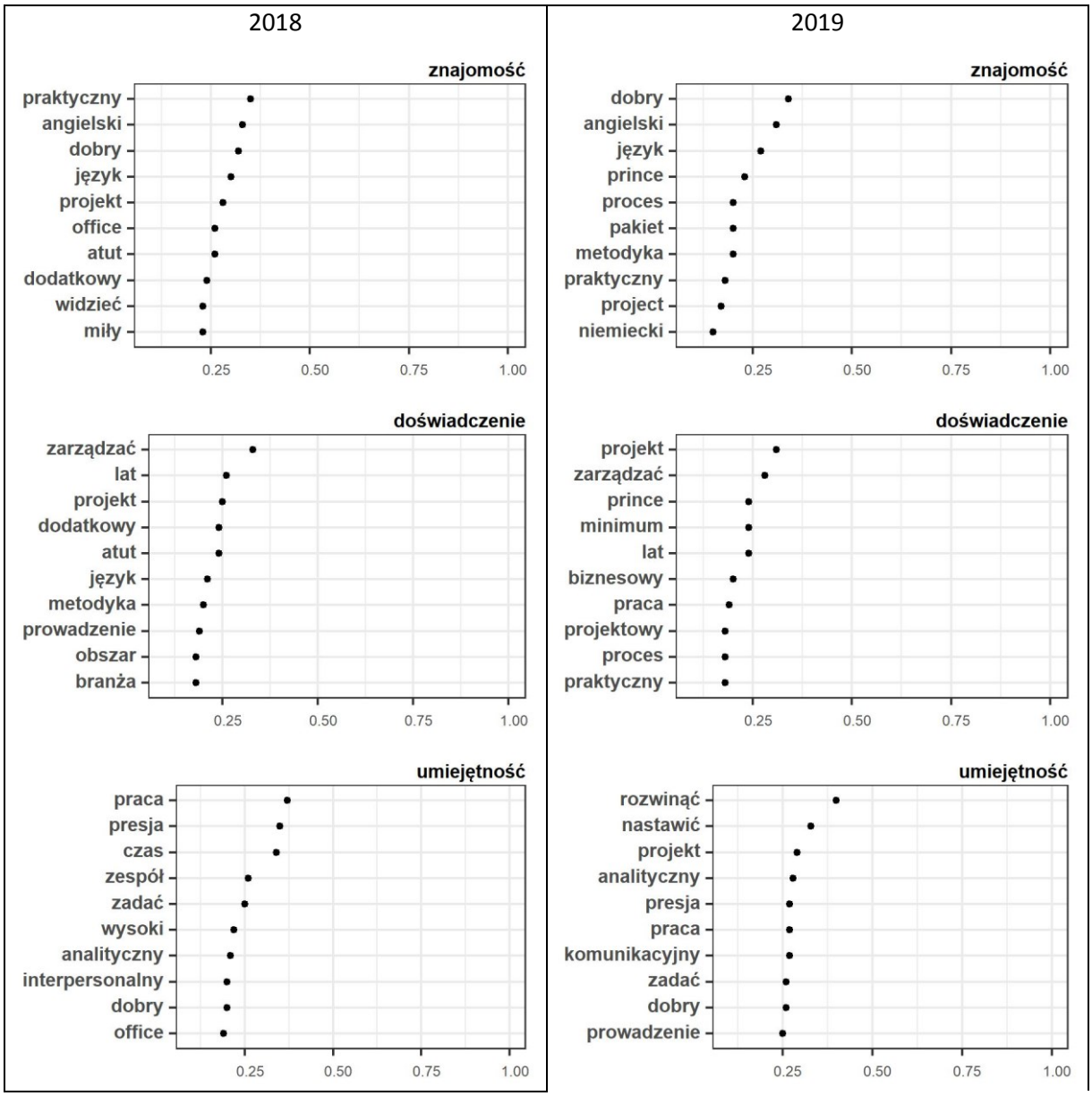

Figure 7. Correlation for selected words used in ads – part one.

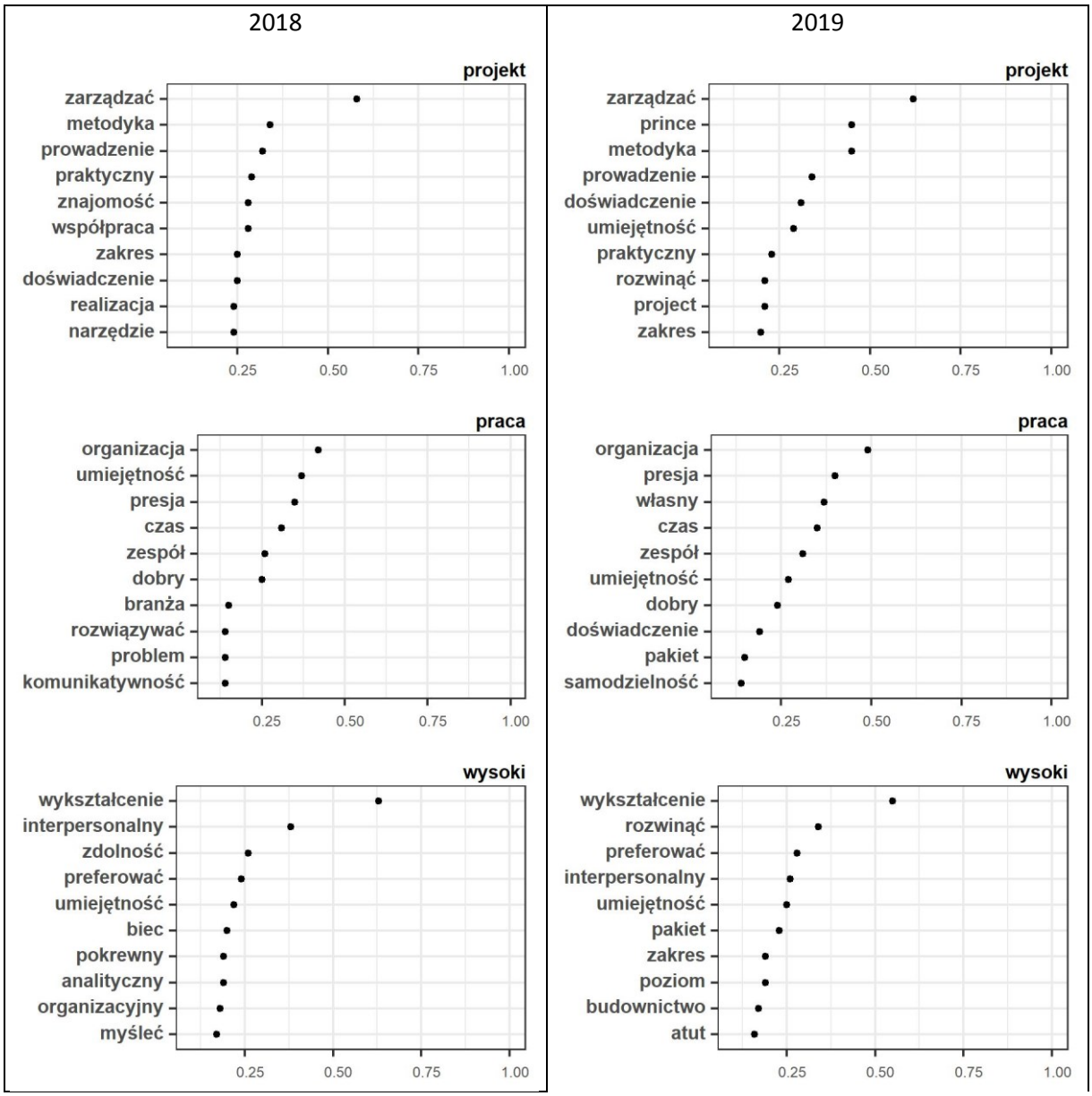

Figure 8. Correlation for selected words used in ads – part two.

Figure 9. Word cloud of identified topics.

Using the Latent Derelicht Allocation (LDA) algorithm, abstract topics describing the analyzed job offers were generated for each corpus. When creating issues, six of the common words were intentionally removed. Words such as „doświadczenie", „praca", „projekt", „umiejętność", „zarządzać", „znajomość" have been removed. The number of generated topics was set by the author at 12. According to the "Griffiths2004" metrics and the "Gibbs" method, the optimal amount of issues for the corpus created by the documents from 2018 was 22.

For the corps comprising the papers from 2019, the number of optimal topics was 25. The generated items were presented in Figure 9 in the form of a word cloud consisting of seven words. On the left are items from 2018 and on the right from 2019. Half of the generated topics are presented. The ones shown were similar (taking into account the most important words). Of course, the issues are not identical, and not all the words that make up the topic are the same—looking through the first pair of items from Fig. 9, it can be seen that these are probably competencies related to construction law (from the words: „prawo", „budowlany") and building permissions (from the terms: „uprawnienia", „budowlany"), as well as permission to drive a car (from the words: „prawo", „jazda", „kat"). The second pair of topics is dominated by the terms „minimum" and „lat".

6. Conclusion

The text mining solution used by the author did not analyze the meaning of words and sentences. It also did not take into account whether the words were next to each other in a sentence. Some information was also lost at the stage of initial text processing, e.g., removing a number, separating names of specific tasks and issues, e.g., „Bezpieczeństwo i higiena pracy", „prawo budowlane", „system zarządzania jakością", „nadzór budowlany". The solution used was to detect certain rules and regularities regarding the occurrence of specific word strings. That is, acquire new previously unknown information, e.g., the most commonly used words and their correlations with other words.

References

1. Gunz, H. (1983). The competent manager: A model for effective performance. R.E. Boyatzis (ed.). New York: Wiley. ISBN 0–471–09031-X. *Strategic Management Journal*, *4(4),* 385–387. https://doi.org/10.1002/smj.4250040413.
2. Hearst, M.A. (1999). *Untangling text data mining*. ACL '99 Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, 3-10. https://doi.org/10.3115/1034678.1034679.
3. Kao, A., & Poteet, S.R. (2007). Natural language processing and text mining. In: A. Kao & S.R. Poteet (Eds.), *Natural Language Processing and Text Mining*. https://doi.org/10.1007/978-1-84628-754-1.
4. Katz, R.L. (1974). Skills of an effective administrator. *Harvard Business Review, vol. 52, no. 5,* 90-102.

5. Le Deist Delamare, F., & Winterton, J. (2005). What Is Competence? *Human Resource Development International*, *8(1),* 27-46. https://doi.org/10.1080/1367886042000338227.
6. Tan, A.-H. (1999). Text Mining: The state of the art and the challenges. *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, *8*, 65-70. https://doi.org/10.1.1.38.7672.
7. Armstrong, M., Wąsik, D., Klimowicz, M., Taylor, S., Patkaniowski, M., Podsiadło, I. (2016). *Zarządzanie zasobami ludzkimi*. Wolters Kluwer Polska.
8. Berry, M.W., & Kogan, J. (2010). Text Mining: Applications and Theory. In *Text*. Wiley.
9. Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation Michael I. Jordan. *Journal of Machine Learning Research, Vol. 3.* Retrieved from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.
10. Gładysz, A. (2012). Zastosowanie metod eksploracyjnej analizy tekstu w logistyce. *Logistyka*, *nr 3*, 643-651. Retrieved from http://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-article-BUS6-0042-0021.
11. Grzegorz, F. (2014). *Zarządzanie kompetencjami. Perspektywa firmowa i osobista*. Retrieved from https://www.empik.com/zarzadzanie-kompetencjami-filipowicz-grzegorz,p1092376800,ksiazka-p.
12. Lula, P. (2005). *Text mining jako narzędzie pozyskiwania informacji z dokumentów tekstowych*. Retrieved from www.statsoft.pl/czytelnia.html67.
13. Mirończuk, M. (2012). Przegląd Metod i Technik Eksploracji Danych Tekstowych. *Studia i Materiały Informatyki Stosowanej*, *4(6),* 25-42. Retrieved from https://repozytorium.ukw.edu.pl/bitstream/handle/item/3527/Przeglad metod i technik eksploracji danych tekstowych.pdf?sequence=1&isAllowed=y.
14. Musioł-Urbańczyk, A. (2010). Kluczowe kompetencje kierownika projektu. *Organizacja i Zarządzanie : Kwartalnik Naukowy*, *nr 2*, 93-108. Retrieved from http://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-article-BSL3-0024-0017.
15. Oleksyn, T. (2006). *Zarządzanie kompetencjami : teoria i praktyka*. Oficyna Ekonomiczna. Oddział Polskich Wydawnictw Profesjonalnych.
16. Pawlak, M. (2006). *Zarządzanie projektami.* Warszawa: PWN.
17. Piwowar-Sulej, K. (2013). Kierownik projektu − charakterystyka profesji. *Nauki Społeczne*, *1(7)*. Retrieved from http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.desklight-d1387a5e-f2ae-4e47-88ae-0a9ed2b54fe7.
18. Spencer, L.M., & Spencer, S.M. (1993). *Competence at work : models for superior performance*. Retrieved from https://archive.org/details/competenceatwork00spen.
19. Wachowiak, P., Gregorczyk, S., Grucza, B., & Ogonek, K. (2004). *Kierowanie zespołem projektowym.* Warszawa: Difin S.A.
20. Wyrozębski, P. (2009). Modele kompetencji w zarządzaniu projektami. *E-Mentor*, *2(29),* 55-64. Retrieved from http://www.e-mentor.edu.pl/artykul/index/numer/29/id/637.