SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2020
ORGANIZATION AND MANAGEMENT SERIES NO. 144

TEXT MINING IN THE IDENTIFICATION OF DUTIES AND RESPONSIBILITIES OF THE PROJECT MANAGER

Marcin WYSKWARSKI

Institute of Economy and Informatics, Faculty of Organization and Management of Silesian University of Technology; marcin.wyskwarski@polsl.pl, ORCID: 0000-0003-2004-330X

Purpose: An attempt to identify the duties and responsibilities of the project manager by analysing job offers from a job website. An attempt to determine whether there were any changes between 2018 and 2019.
Design/methodology/approach: Text mining was performed for fragments of job offers, describing the duties and responsibilities. The text mining analysis consisted of initial processing of the text, creation of a corpus of analysed documents, construction of a word frequency matrix and use of classical methods from the data mining are.
Findings: The most common words in job offers are presented, as well as their correlation with other words. With the use of the Topic modeling algorithm, hidden topics describing the analysed job offers have been generated. These topics can also be used to identify the duties and responsibilities of a project manager.
Research limitations/implications: Only the job offers meeting the following conditions were analysed: (1) they concerned the job of „project manager"; (2) the content was in Polish; (3) they were provided by www.pracuj.pl website; (4) they were collected from 09 to 11 April in 2018 and 2019.
Practical implications: This method can be used by organizations training project managers, in order to modify and better adjust the curriculum to the needs of the labour market.
Originality/value: Research has shown that text mining can be used to determine the responsibilities of a project manager by analysing job offers.

Keywords: text mining, duties and responsibilities, project manager, word cloud, topic modeling.

Category of the paper: Research paper, case study.

1. Introduction

The source literature on project management and general management provides information on the basic duties of a project manager, tasks performed and problems encountered (Wachowiak et al., 2004; Walczak, 2014; Pawlak, 2006; Steyn and Nicholas, 2012; Trocki,

2013). Obtaining information about the duties and responsibilities of a project manager may be useful for persons interested in working at this position, as well as for organizations training project managers, in order to modify and better adjust the curriculum.

The aim of this article was to identify the duties and responsibilities of a project manager and to determine whether changes in this area have occurred over time. For this purpose, the text mining of job offers from a job website was analysed.

2. Data source and text mining analysis

Text analytics were performed for the job offers available at www.pracuj.pl website. The data was collected from 09 to 11 April 2018 and 09 to 11 April 2019. In order to find offers, the phrase "project manager" (pl: „kierownik projektu") was used. Among the offers provided by www.pracuj.pl, there were also those, that used a different name for the position than „project manager" (e.g. „projects' manager", „project supervisor", „program manager", „program coordinator"). Some of the offers were in English. Only the offers meeting the following two conditions were further analysed:

- the offer concerned the job of „project manager" (pl: „kierownik projektu") (e.g. offers for the job of „project coordinator", „program manager" etc. were not analysed),
- the content of the offer was in Polish.

Out of the 775 offers found in April 2018, 369 were selected for analysis. Out of the 707 offers found in April 2019, 350 were selected for analysis. The number of job offers (all of them and those selected for analysis) from particular voivodeships is shown in Figure 1 and 2.

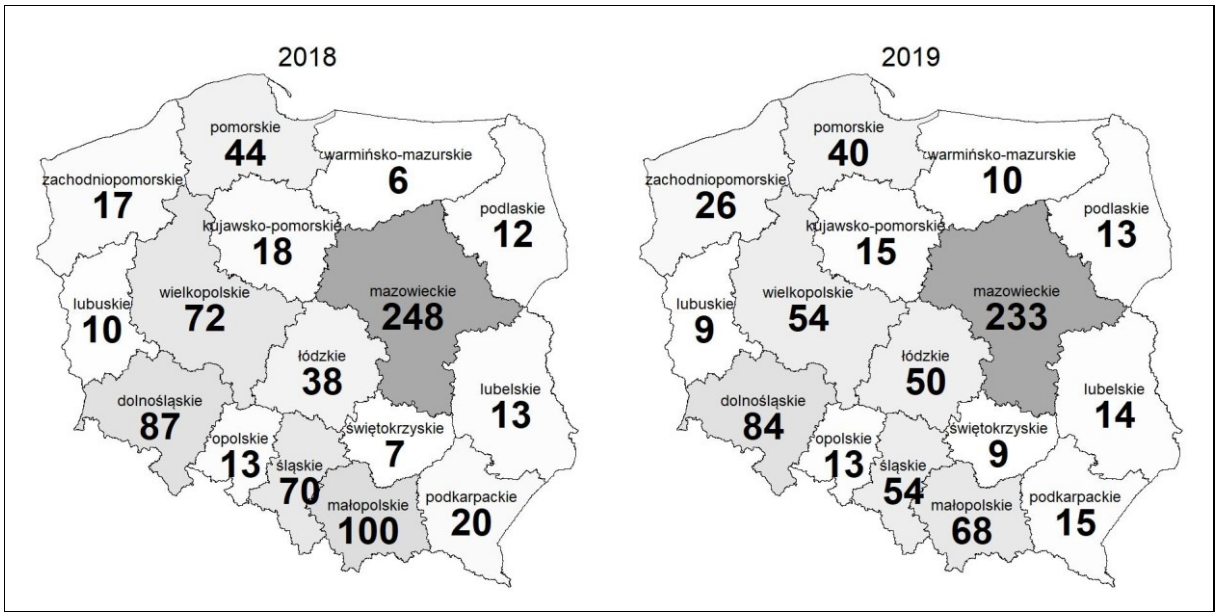

Figure 1. All job offers, own elaboration.

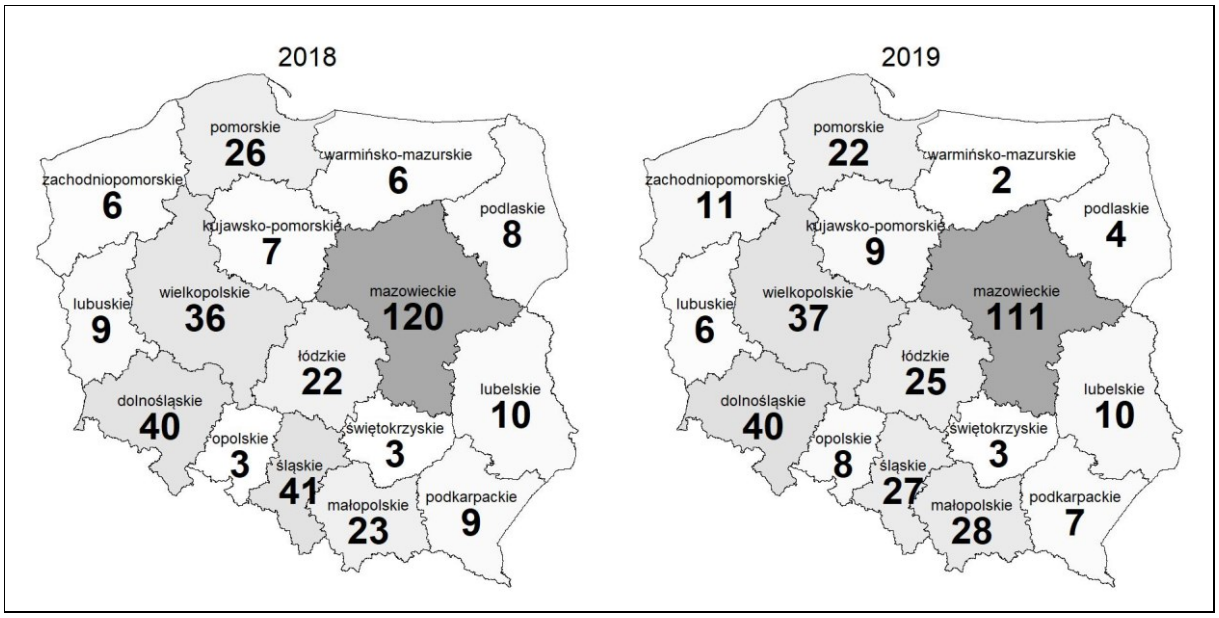

Figure 2. Analysed job offers, own elaboration.

For each offer approved for analysis, a file with the .txt extension was created, in which a fragment describing the duties and responsibilities was saved. This part of the offer was defined by different phrases. The most popular (for 2018 and 2019 in total) are shown in Figure 3. The graph shows phrases that have appeared at least 8 times. In order to ensure the possibility of analysing offers for a selected voivodeship in a given year, the created files were placed in separate folders, divided into voivodeships and the year of downloading the offer (32 folders).

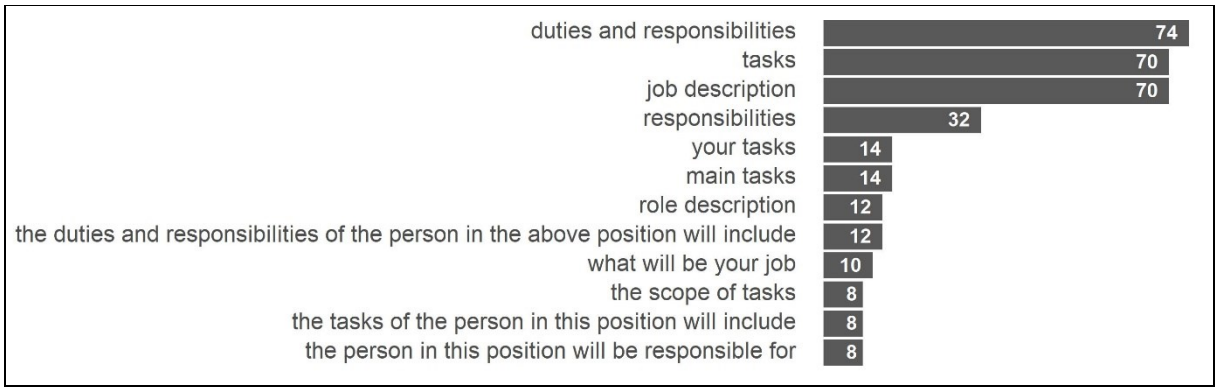

Figure 3. Frequently used terms for the "duties and responsibilities" of a project manager, own elaboration.

Figure 4 shows the arrangement of the number of words in the created text files for 2018 and 2019 offers in the form of a histogram and a box diagram. It is the number of words after all characters, except for the letters, have been deleted from the files. It can be observed that these are relatively short documents (the shortest ones consisted of only 8 words and the average number of words is 50).

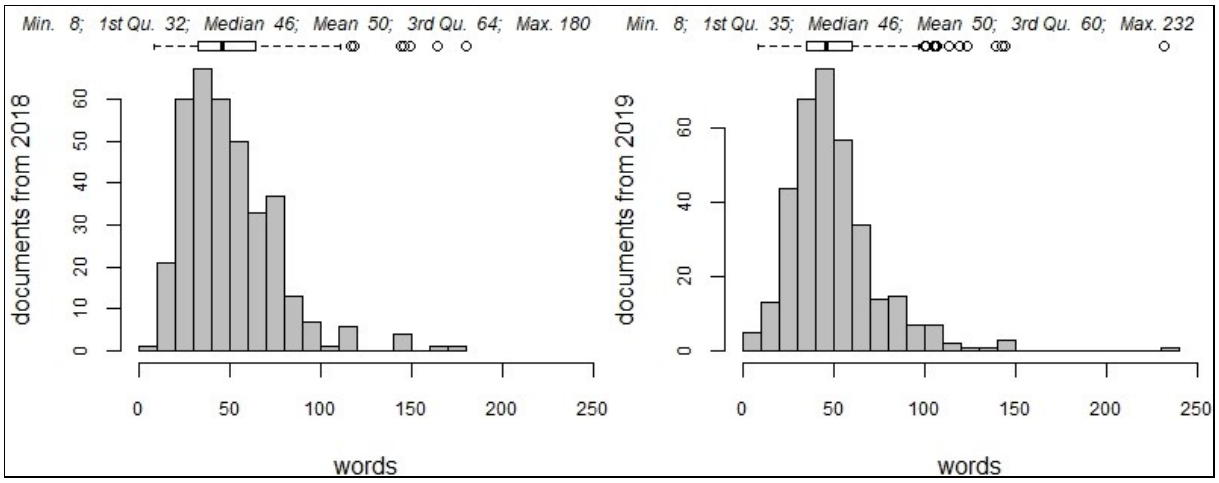

Figure 4. Number of words in created text documents, own elaboration.

The text mining analysis consisted of the following steps:

- initial processing of the text,
- creation of a corpus of analysed documents, construction of a word frequency matrix,
- use of classical methods from the data mining area.

The result of the initial processing of the text was the transformation of each text file into a so-called word bag. For this purpose, by using Notepad++ v.7.3.3 and RStudio v.1.0.136 applications, the following actions were performed:

- all characters except for the letters have been deleted,
- the capital letters were changed to lower case,
- words considered useless (e.g. conjunctions, prepositions etc.) were deleted – for this purpose a stopword for Polish language was used,
- words have been converted to their basic version,

each word was placed in a separate line.

The process of transforming words into their basic form was carried out using a morphosyntactic dictionary of the Polish language "polimorfologik 2.1". This dictionary, in the form of a text file, was downloaded from the Github website[1]. At the time of using the dictionary, it contained 4,811,854 lines of text. After importing the dictionary into the RStudio program, it became a table consisting of three columns: basic form, modified form, grammatical markers. The transformation of a word into a basic form consisted in finding it in the "modified form" column and replacing it with the "basic form" column. If the word was not found in the dictionary, it remained unchanged in the document.

The next step in text mining analysis was to create two document corpuses. The first corpus was composed of offers collected in 2018 and the second one from 2019. In the next step, a document term matrix with a Term Frequency (TF) representation was created for each corpus. Vector Space Model (Gładysz, 2012; Mirończuk, 2012) was used to create the matrix.

[1] https://github.com/morfologik/polimorfologik/releases/tag/2.1.

A part of the matrix for the corpus made up of offers collected in 2019 is shown in Figure 5 (the original matrix is 350 x 82, i.e. 350 documents and 82 words).

	Terms				
Docs	przygotowywać	przygotowywanie	raportować	raportowanie	realizacja
Z_20.txt	0	0	0	0	4
Z_21.txt	0	0	0	0	1
Z_22.txt	0	0	0	0	0
Z_23.txt	0	0	1	0	0
Z_24.txt	0	0	0	0	1

Figure 5. Extract from the document matrix – expressions for the 2019 Corpus, own elaboration.

In the last step of the analysis, the most common words were searched for the created corpus, which are presented in a bar chart. At this stage, the correlation between the selected four words and other words was also calculated. This correlation was calculated using the findAssocs() function, which is based on the standard cor() function available in the R language statistical package. A correlation of 1 means that the two words always appear together in documents (in the same number). A 0 value means that words never occurred together in the analysed documents.

In the last part of the analysis, the Latent Dirichlet Allocation (LDA) algorithm was used, which is a popular Topic Modeling algorithm. The algorithm was used to generate abstract, hidden topics, describing the analysed job offers. This algorithm assumes that each document is represented by a division into topics, and that each topic is represented by a division into words. The identified topics were intended to facilitate the identification of the duties and responsibilities of a project manager, enable their possible grouping (e.g. team management responsibilities, construction project responsibilities etc.) and enable the comparison of the results for the offers from both periods.

3. Results of the text mining analysis

The results of the text mining analysis are presented in graphic form in Figures 6 to 10.

In Figure 6, in the form of a bar chart, the fifty most frequently used words (together with the number of their appearances) in the 2018 and 2019 offers are shown. The most frequently used word, i.e. „projekt” (en: „project”), was deliberately removed from the chart. The grey colour was used to mark words (7 words in each list) that were only included in one of the lists For example, the word „finansowy” (en: „financial”) was in the fifty most frequently used words only in 2018 offers. As you can see, the order of the three most frequently used words is the same for 2018 and 2019 offers. On the basis of the presented lists, it is possible to determine what tasks the project manager will have to deal with, e.g. „project team management”. (words - pl: „zarządzać”, „zespół”, „projektowy”; en: „manage”, „team”, „project”), „keeping records”

(words - pl: „prowadzić", „dokumentacja"; en: "keep", "records"), „preparation of a schedule" (words - pl: „przygotowywać", „harmonogram; en: „prepare", „schedule").

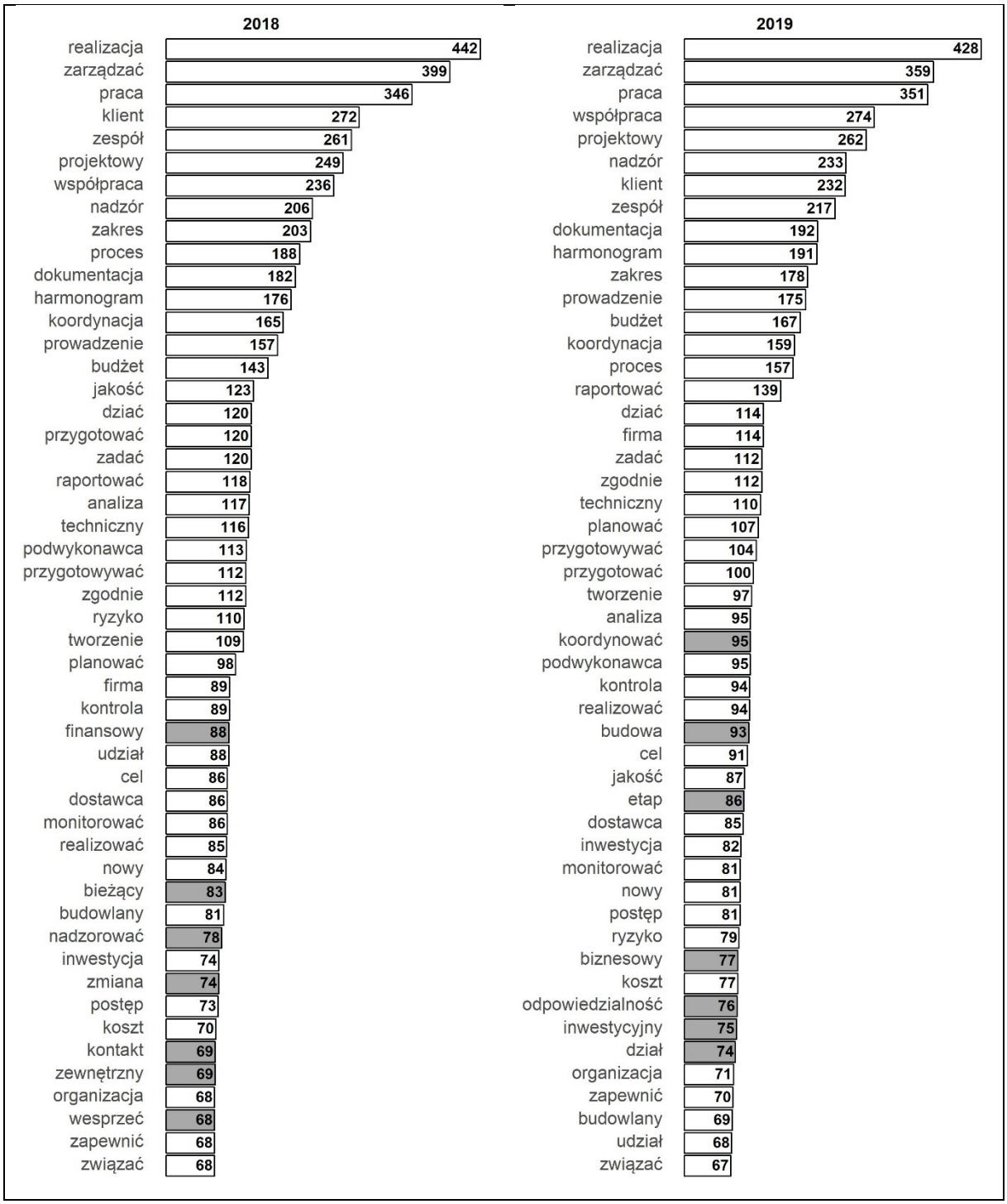

Figure 6. The fifty most frequently used words in the offers, own elaboration.

Figure 7 shows the correlation between particular words and other words in the offers. A 0 value means that words never occurred together. A correlation of 1 means that the two words always appear together (in the same number) in the processed documents. The correlation is presented in the form of scatter graphs divided into 2018 and 2019. Due to the volume, the correlation of only four words selected by the author is presented.

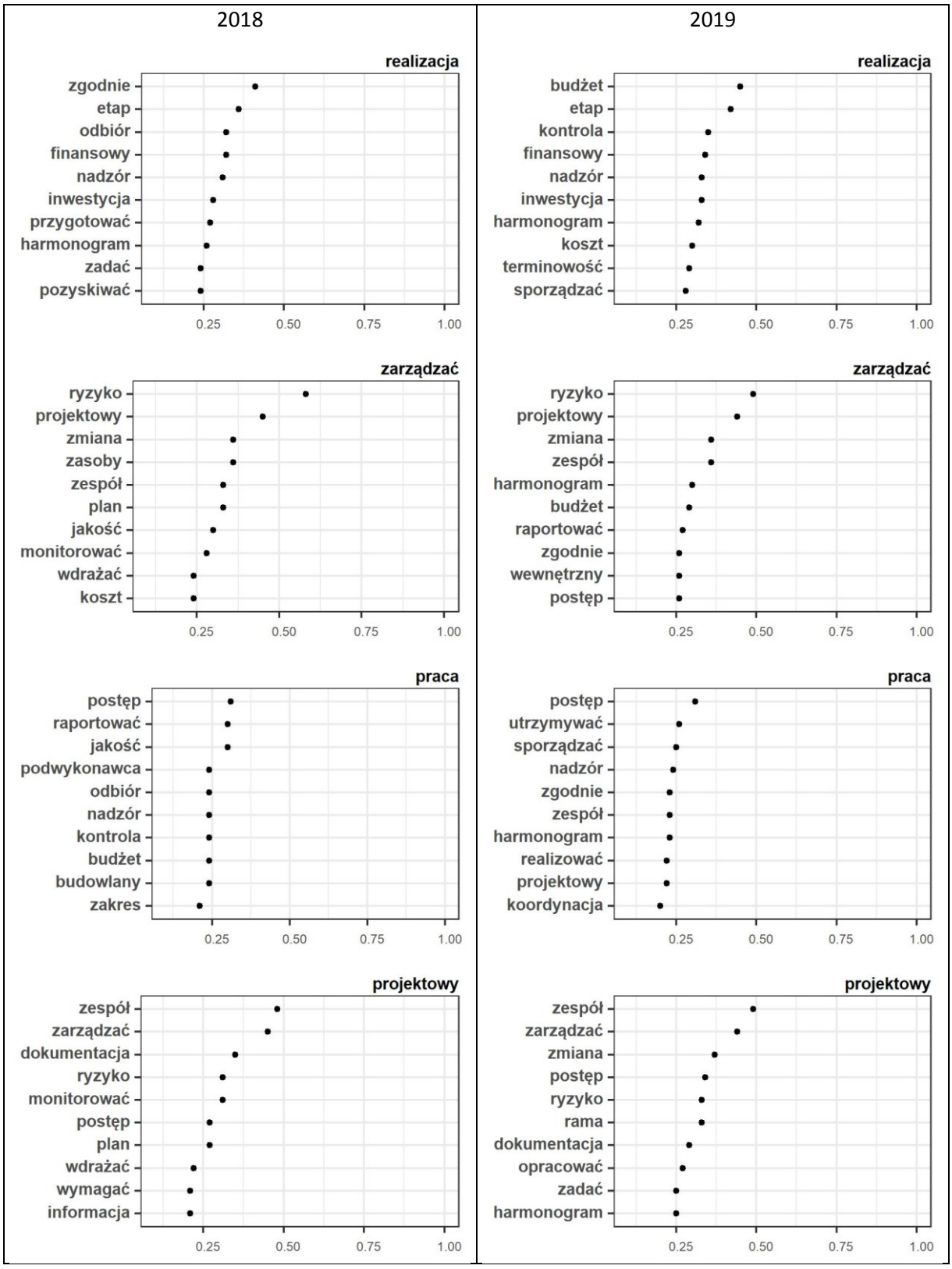

Figure 7. Correlation for selected four words used in the offers, own elaboration.

By reviewing the graphs with correlation values, it is possible to try to establish the tasks of the project manager and to assess whether there have been changes over time. For example, a graph showing the correlation for the word „zarządzać" (en: "manage") can be used to create tasks such as: "risk management" (words – pl: „zarządzać", „ryzyko"; en: "manage", "risk"),

“team management” (words – pl: „zarządzać”, „zespół”; en: “manage”, “team”). From the correlation to the word „praca” (en: „work”) for the year 2018, the task “reporting on the work progress” (words – pl: „raportować”, „postęp”, „praca”; en: “report”, “progress”, “work”) and for 2019 such as: “execution of works in accordance with the schedule” (words – pl: „praca”, „zgodnie”, realizować”, „harmonogram”; en: “work”, “accordance”, “execution”, “schedule”).

Using the Latent Dirichlet Allocation (LDA) algorithm, abstract topics were generated for each corpus, describing the analysed job offers. The most frequently used word, i.e. the word “project”, was deliberately removed. The author determined that the number of topics was 12. The optimal number of topics for both corpuses, according to the “Griffiths2004” metric and the “Gibbs” method, was 27. The topics are presented in the form of a cloud of seven words. They are shown in Figures 8 and 9. The topics from 2018 are on the left and from 2019 on the right. The order of the topics in the figures is based on the size of the words, to make it easier to compare the topics from both years.

The topics generated by the Latent Dirichlet Allocation algorithm can also be used to identify the duties and responsibilities of a project manager and assess whether there have been significant changes over time. It can be noticed (taking into account the longest word of the word cloud) that nine topics from 2018 and 2019 are similar to each other (the whole Figure 8 and the upper half of Figure 9). Obviously, topics are not identical and not all words included in a given topic are the same. Looking at the first couple of topics in Figure 8, it can be noticed that these are, most likely, activities related to risk management (words – pl: „zarządzać”, „ryzyko”; en: „manage”, „risk”), from the next pair a task related to “project team management” can be created (words – pl: „zarządzać”, „zespół”, „projektowy”; en: „manage”, „team”, „project”). Looking at the second pair of topics in Figure 9, it can be seen that the tasks of the project manager will concern the customer (word – pl: „klient”; en: „customer”). From the 2018 topic it is possible to identify the tasks related to “acquiring a new customer” (words pl: „pozyskiwać”, „nowy”, „klient”; en: „acquire”, „new”, „customer”), „maintaining the relationship with the customer” (words pl: „utrzymywać”, „relacja”, „klient”; en: „maintain”, „relationship”, „customer”). From the topic for 2019, the task of “preparing offers for the customer” can be created (words – pl: „przygotowywać”, „oferta”, „klient”).; en: „prepare”, „offer”, „customer”).

Figure 8. Word clouds of identified topics – part one, own elaboration.

Figure 9. Word clouds of identified topics – part two, own elaboration.

4. Summary

The text mining solution used by the author did not analyse the meaning of words and sentences. It also did not take into account the fact, whether given words appeared next to each other in a sentence. Some information was also lost at the stage of initial text processing, e.g. by deletion of a digit, separation of names of specific tasks and issues, e.g. “Health and Safety at Work”, “Construction Law”, “Quality Management System”, “Construction Supervision”. The applied solution was supposed to detect certain rules and regularities concerning the occurrence of specific sequences of words, that is, to obtain new, previously unknown information, e.g. the most frequently used words and their correlation with other words.

References

1. Gładysz, A. (2012). Zastosowanie metod eksploracyjnej analizy tekstu w logistyce. *Logistyka, 3,* pp. 643-651.
2. Mirończuk, M. (2012). Przegląd Metod i Technik Eksploracji Danych Tekstowych. *Studia i Materiały Informatyki Stosowanej, 4(6),* pp. 25-42.
3. Pawlak, M. (2006). Zarządzanie *projektami*. Warszawa: PWN.
4. Steyn, H., and Nicholas, J.M. (2012). *Zarządzanie projektami. Zastosowanie w biznesie, inżynierii i nowoczesnych technologiach.* Warszawa: Wolters Kluwer.
5. Trocki, M. (2013). *Nowoczesne zarządzanie projektami*. Warszawa: PWE.
6. Wachowiak, P. et al., (2004). *Kierowanie zespołem projektowym.* Warszawa: Difin.
7. Walczak, R. (2014). *Podstawy zarządzania projektami. Metody i przykłady*. Warszawa: Difin.

Appendix – translation of words used in figures

analiza - analysis
bieżący - current
biznesowy - business
budowa - building
budowa - construction
budować - build
budowlany - construction
budżet - budget
cel - target
dać - give
definiować - define
dokumentacja - records
dostawca - supplier
dotyczyć - regarding
działać - act
dział - department
etap - stage
finansowy - financial
firma - enterprise
harmonogram - schedule
informacja - information
inwestycja - investment
inwestycyjny - investment
jakość - quality
kierować - manage
kierownik - manager
klient - customer
komunikacja - communication
kontakt - contact
kontrola - verification
koordynacja - coordination
koordynować - coordinate
koszt - cost
monitorować - monitor
nadzorować - supervise
nadzór - supervision
niezbędny - essential
nowy - new
obszar - area
odbiór - acceptance
odpowiedzialność - responsibility
oferta - offer
opracować - develop
organizacja - organization
plan - plan
planować - plan
podwykonawca - subcontractor
postęp - advancement
poszczególny - specific
powierzyć - entrust
pozyskiwać - obtain
praca - work
proces - process
produkcja - production
produkt - product
projektowy - project
prowadzenie - operating
prowadzenie - running
prowadzić - operate
przygotować - prepare
przygotowywać - prepare
przygotowywanie - preparation
rama - frame
raport - report
raportować - report
raportowanie - reporting
realizacja - execution
realizować - execute
relacja - relation
robota - labor
rozliczać - settle
rozwój - development
ryzyko - risk
sporządzać - draft
system - system
techniczny - technical
terminowość - punctuality
tworzenie - creation
udział - involvement
udział - participation
utrzymywać - maintain
wdrażać - implement
wdrożenie - implement
weryfikacja - reviewing
wesprzeć - promote
wesprzeć - support
wewnętrzny - internal
współpraca - cooperation
wykonać - execute
wykonawca - contractor
wymagać - require
zaangażować - engage
zadać - assign
zakres - scope
zapewnić - provide
zarządzać - manage
zasoby - resources
zespół - team
zewnętrzny - external
zgodnie - consistent
zmiana - change
związać - bin