ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2017
Seria: ORGANIZACJA I ZARZĄDZANIE z. 114 Nr kol. 1993

Marcin WYSKWARSKI
Politechnika Śląska
Wydział Organizacji i Zarządzania
marcin.wyskwarski@polsl.pl

TEXT MINING W ANALIZIE ZBIORÓW PUBLIKACJI NAUKOWYCH

Streszczenie. Często stosowaną form przechowywania informacji w organizacjach i społeczeństwie jest tekst. Tekst może zostać poddany eksploracji w celu pozyskania wcześniej nieznanej i użytecznej wiedzy. Celem niniejszego artykułu jest prezentacja wyników analizy wybranych publikacji naukowych. Analiza została przeprowadzona z wykorzystaniem text mining'u, a jej celem było ustalenie jakich słów najczęściej używali autorzy publikacji, oraz obliczenie korelacji tych słów z innymi.

Słowa kluczowe: text mining, chmura słów, publikacje naukowe

TEXT MINING IN ANALYSIS OF SCIENTIFIC PUBLICATIONS

Abstract. Text is often used to store information in organizations and society. Text can be explored to gain previously unknown and useful knowledge. The aim of this article is to present the results of the analysis of selected scientific publications. The analysis was done using text mining. Its purpose was to determine what words were most used by the authors, and to calculate the correlation of those words with others words.

Keywords: text mining, word clouds, text mining, scientific publications

1. Wstęp

Jedną z cech współczesnego świata są stale i szybko rosnące zasoby informacji. Często problemem nie jest już brak informacji lecz jej nadmiar. Wzrasta liczba zasobów informacyjnych organizacji gospodarczych, publicznych, społecznych ale również osób prywatnych. W obszarze nauki również zauważa się dynamiczny przyrost informacji gromadzonych w postaci publikacji naukowych. Fakt ten sprawia, że zapoznanie się z bieżącą

literaturą dotycząca wybranego obszaru może okazać się bardzo trudne, lub wręcz niemożliwe.

Jednym ze sposobów wspomagania analizy informacji przechowywanych w formie tekstu pisanego jest zastosowanie odpowiednich rozwiązań informatycznych, które pozwalają całkowicie, lub częściowo zautomatyzować proces przetwarzania i analizowania tekstu. Jednym z takich rozwiązań jest tzw. eksploracyjna analiza tekstu (ang. text mining), dzięki której można przyśpieszyć proces wydobywania informacji z zasobów tekstowych.

Celem niniejszego artykułu jest prezentacja wyników analizy wybranych publikacji naukowych. Analiza została przeprowadzona z wykorzystaniem text mining'u, a jej celem było wyszukanie najczęściej używanych słów, oraz ustalenie ich związku z innymi słowami. Adekwatnie do przyjętego celu podporządkowana została struktura pracy. W punkcie drugim przedstawiono podstawowe informacje związane z przeprowadzaniem analizy tekstu z wykorzystaniem text mining'u. Punkt trzeci omawia sposób przeprowadzonej analizy, oraz prezentuje uzyskane wyniki.

2. Text mining - istota

Text mining to stosunkowo młoda i interdyscyplinarna dziedzina. Wywodzi się m.in. z data mining, wyszukiwania informacji, kategoryzacji tekstu i modelowania probabilistycznego [4]. Według Marti Hearst text mining to "proces mający na celu wydobycie z zasobów tekstowych nieznanych wcześniej informacji" [3]. To metody, koncepcje oraz algorytmy przetwarzania zasobów tekstowych sporządzonych w językach naturalnych, które są implementowane w postaci programów komputerowych, co prowadzi do zautomatyzowania procesów przetwarzania dokumentów [2].

Zakres zastosowań text miningu jest dość rozległy. Według P. Lula najpopularniejsze obszary to pozyskiwanie informacji z dokumentów, identyfikacja wiadomości zawierających określone treści, generowanie streszczeń , klasyfikacja wzorcowa, klasyfikacja bezwzorcowa (grupowanie, klasteryzacja), identyfikacja powiązań, wizualizacja oraz generowanie odpowiedzi na pytanie [5].

A. Tan podaje, że 80% informacji przedsiębiorstw jest przechowywanych w dokumentach tekstowych, co przyczynia się do ciągłego zainteresowania tą tematyką [7]. Na popularność text miningu wpływa również nieustanny rozwój Internetu i towarzyszący mu wzrost znaczenia opiniotwórczej roli portali społecznościowych [1].

Duża część rozwiązań text miningowych nie analizuje znaczenia wyrazów oraz zdań, lecz próbuje wykryć reguły i prawidłowości dotyczące występowania określonych ciągów znaków – czyli słów. W dużym uproszczeniu proces analizy tekstu można podzielić na trzy części tj.

wstępne przetwarzanie tekstu, budowę macierzy częstości występowania słów oraz wykorzystanie klasycznych metod pochodzących z obszaru data mining [2].

Wstępne przetwarzanie tekstu polega na przekształceniu dokumentu tekstowego (dokumentów tekstowych) w listę wyrazów określanych mianem worka słów (ang. bag of words). Pomijane są informacje o kolejności słów i związkach pomiędzy nimi, przyjmuje się, że wystąpienia słów są niezależne od siebie [2].

Z dokumentu tekstowego usunięte zostają cyfry, oraz wszelkie znaki interpunkcyjne (np. kropki, przecinki, średniki, myślnik). W kolejnym kroku bez szkody dla jakości analizy, dzięki tzw. stop-liście (ang. stop-words) usunięte zostają słowa nie wnoszące dodatkowych informacji (usunięte zostają m.in. spójniki, przyimki itp.). Usunięciu mogę ulec również inne nieprzydatne w analizie słowa, tzn. występujące bardzo rzadko (ang. least frequent) bądź bardzo często (ang. most frequent) - tzw. przycinanie (ang. pruning).

W celu przekształcenia różnych form tego samego słowa do wersji uznawanej za podstawową stosuje się tzw. ekstrakcję rdzeni słów (ang. stemming) oraz lematyzację. Stemming polega na wybraniu z danego słowa, części niezmiennej dla wszystkich form gramatycznych czyli tzw. rdzenia (to usunięcie wszelkiego rodzaju przedrostków i przyrostków). Lematyzacja to natomiast analiza morfologiczna umożliwiająca znalezienie podstawowej formy danego wyrazu tj. identyfikacji leksemu (np. czasownik zostaje przekształcony do bezokolicznika, a rzeczownik do mianownika liczby pojedynczej). Do przeprowadzenia lematyzacji niezbędny jest słownik lub rozbudowany zestaw reguł fleksyjnych dla danego języka.

Inna możliwą do wykonania operacją podczas wstępnego przetwarzania tekstu jest tagowanie (ang. tagging), które polega na wyborze odpowiedniego opisu morfoskładniowego [6].

Budowa macierzy częstości występowania słów wykorzystuje tzw. model przestrzeni wektorowej (ang. Vector Space Model), w którym dokumenty i występujące w nich słowa są reprezentowane w postaci macierzy. Każdy analizowany dokument jest reprezentowany przez osobny wektor przedstawiający liczbę wystąpień poszczególnych słów [2, 6]. Macierzową postać reprezentacji dokumentów oraz związanych z nimi słów określa się macierzą dokumentów-wyrażeń (ang. document - term matrix). Jej wiersze reprezentują analizowane dokumenty, a kolumny znajdujące się w nich słowa[1]. W zależności od przyjętego sposobu kodowania informacji w elementach macierzy, istniej możliwość uzyskania różnych odmian reprezentacji przestrzenno-wektorowej tekstu. Wśród często stosowanych wymienia się reprezentację boolowska (binarna), częstotliwościową występowania wyrażeń (ang. term frequency - TF), odwrotnej częstości dokumentu (ang. inverse-documentfrequency – IDF), mieszaną TF-IDF, logarytmiczną, ważoną logatytmiczną, okapi BM25 [6]. W lingwistyce informatycznej spotyka się także określenie korpus dokumentów. Korpus jest kolekcją

[1] Utworzona może zostać także tzw. macierz wyrażeń-dokumentów (ang. term - document matrix), w której dokumenty są prezentowane przez kolumny, a słowa przez wiersze.

dokumentów, które będą analizowane. Na jego podstawie tworzona jest macierz dokumentów-wyrażeń lub macierz wyrażeń-dokumentów.

Po wstępnym przetwarzaniu tekstu i utworzeniu macierzy częstości występowania słów, można przejść do kolejnych czynności związanych z analizowanym tekstem np. do grupowania dokumentów tekstowych (tzw. klasteryzacji), obliczenia częstości występowania słów, korelacji między używanymi słowami, wizualizacji przeprowadzonych obliczeń itd.

3. Analiza text minig wybranych zeszytów naukowych serii Organizacja i Zarządzanie

Analizie poddano zawartość trzech zeszytów naukowych serii „Organizacja i Zarządzanie". Zeszyty zostały opublikowane w 2016 r. i miały zbliżoną liczbę artykułów. Przeanalizowano tylko artykuły, których językiem podstawowym był język polski.

Tabela 1

Informacja o zeszytach naukowych poddanych analizie
Information about the scientific papers analyzed

Nr zeszytu naukowego	Liczba artykułów w zeszycie	Liczba artykułów wybranych do analizy
89	42	42
95	44	44
97	41	38

Źródło: Opracowanie własne.

W opracowaniu przeprowadzono podstawową analizę tekstu, której celem było:

- wyszukanie najczęściej używanych słów - przedstawiono je w postaci wykresu słupkowego prezentującego liczbę użytych słów, oraz tzw. chmury słów,
- określenie związku najczęściej używanych słów z innymi słowami (na bazie współczynnika korelacji).

Z artykułów utworzono cztery oddzielnie analizowane korpusy. Korpus pierwszy stanowiły artykuły z zeszytu nr 89, korpus drugi z zeszytu 95 a korpus trzeci z zeszytu 97. Korpus czwarty objął artykuły wchodzące w skład korpusów 1, 2 i 3. Liczbę artykułów poddanych analizie przedstawia tabela 1. Pominięto trzy artykuły z zeszytu nr 97, gdyż zostały one opublikowane w języku angielskim.

W przeprowadzonej analizie zostały wykorzystane następujące aplikacje:

- JDownloader - niezależna, open source'owa platformowa programowa do pobierania plików z Internetu;

- Free PDF to Text Converter - program firmy Free PDF Solutions Inc przeznaczony do zmiany formatu plików,
- Notepad++ v.7.3.3 - edytor tekstu,
- RStudio v.1.0.136 - zintegrowane środowisko programistyczne dla języka R.

Aplikacja „JDownloader" została wykorzystana do automatycznego pobrania plików pdf z artykułami, które zostały zamieszone na stronie własnej Zeszytów Naukowych Politechniki Śląskiej Seria Organizacja i Zarządzanie[2].

Za pomocą aplikacji "Free PDF to Text" zmieniono format plików artykułów z pdf na txt. Format plików został zmieniony z dwóch powodów:

- problemów z poprawnym odkodowaniem polskich znaków podczas tworzenia korpusów w aplikacji RStudio na podstawie plików pdf,
- uzyskania możliwości oczyszczania dokumentów ze zbędnych nieprzydatnych fragmentów.

Do wstępnego oczyszczenia dokumentów posłużyła aplikacja Notepad++, w której skorzystano z mechanizmu wyrażeń regularnych (ang. regular expressions) i możliwości tworzenia makr automatyzujących ten proces. Za pomocą makr połączono w całość słowa, które były w dwóch częściach w wyniku wcześniejszego zastosowania podziału wyrazów, oraz usunięto:

- zawartość nagłówka i stopki,
- tytuł, streszczenia i słowa kluczowe napisane w języku angielskim,
- spis literatury,
- dane osobowe autora/autorów oraz afiliację,
- słowa uznane w tej analizie przez autora za tzw. sotp word's:
  - m.in. słowa „wstęp", „zakończenie", „streszczenie", „podsumowanie" stanowiące tytuły rozdziałów,
  - słowa „tabela", „rysunek", „tab.", „rys" (oraz ich odpowiedniki w języku angielskim), które znajdowały się w tytułach tabel i rysunków,
  - słowo „źródło" znajdujące się pod rysunkiem / tabelą[3].

Dalsze operacje czyszczenia dokumentów przeprowadzono z wykorzystaniem aplikacji RStudio. Za pomocą wyrażeń regularnych usunięto wszystkie znaki z wyjątkiem liter, zamieniono duże litery na małe, a następnie podzielono zawartość dokumentu w taki sposób, aby każdy wyraz znajdował się w osobnej linii.

Kolejne działania również zostały przeprowadzone w aplikacji RStudio. Istotnym zagadnieniem było sprowadzenie słów do ich podstawowej formy oraz usunięcie tzw. stop-word's. Niesyty funkcje pakietu „TM" (opracowanego do przeprowadzania analizy Text Mining w języku R) nie zapewniają operacji stemming'u, lematyzacji oraz usunięcia stop-

[2] adres www strony to: http://organizacjaizarzadzanie.blogspot.com

[3] nie było usuwane słowo „źródło" znajdujące się w treści danego artykuł, np. ze zdania „Opinie klientów są istotnym źródłem informacji dla przedsiębiorstwa".

word's dla języka polskiego. W związku z powyższym w celu przekształcenia słów do ich formy podstawowej wykorzystano słownik morfosyntaktyczny „polimorfologik 2.1" udostępniony na portalu Github[4]. Słownik ten stanowi zwykły plik tekstowy w kodowaniu UTF-8 o formacie: forma podstawowa; forma odmieniona; znaczniki gramatyczne. Po prawidłowym zaimportowaniu pliku do programu RStudio zawartość pliku może przyjąć postać tabeli składającej się trzech kolumn. W momencie korzystania ze słownika miał on 4 811 854 linii tekstu[5]. W tabeli 2 przedstawiono przykładowe trzy wyrazy z tego pliku (zestaw użytych znaczników gramatycznych został przez autorów opisany w dokumentacji).

Tabela 2

Struktura pliku polimorfologik 2.1

Forma podstawowa	Forma odmieniona	Znaczniki gramatyczne
projektowy	projektowo	adja
projekt	projektu	subst:sg:gen:m3
projektujący	projektującemu	adj:sg:dat:m1.m2.m3.n1.n2:pos+subst:sg:dat:m1

Źródło: Opracowanie własne.

Przekształcenie słowa do postaci podstawowej polegało na odszukaniu go w kolumnie „Forma odmieniona" i jego zamianie na słowo znajdujące się w tym samym wierszu, w kolumnie „Forma podstawowa". Dzięki znacznikom gramatycznym do dalszej analizy zostały wybrane rzeczowniki oraz przymiotniki. Jeżeli dane słowo nie zostało odnalezione w kolumnie „Forma odmieniona" to do dalszej analizy przechodziło w niezmienionej formie (dlatego występujące w tekście słowa w języku innym niż polski nie zostały usunięte) - w tym przypadku nie uwzględniano także formy gramatycznej danego słowa.

W celu usunięcia tzw. stop-word's utworzono własną listę tych słów dla języka polskiego składającą się z 407 słów. Wykorzystano listy zamieszczone w Internecie[6] oraz słowa wytypowane przez autora tej analizy (np. m.in. liczebniki – pierwszy, pierwszym, pierwszego itd.).

	Terms			
Docs	inspekcja	inspektor	inspiracja	inspirowanie
89-37.txt	4	2	0	1
89-38.txt	0	0	0	0

Rys. 1. Fragment macierzy dokumentów-wyrażeń dla korpusu 1
Źródło: Opracowanie własne.

Kolejnym krokiem było utworzenie czterech korpusów, dla których wykonano macierze dokumentów-wyrażeń (ang. document - term matrix) z częstotliwościową reprezentacją występowania wyrażeń (ang. term frequency - TF). Na rysunku 1 przedstawiono fragment

[4] https://github.com/morfologik/polimorfologik/releases/tag/2.1
[5] maj 2017.
[6] np. lista ze strony http://www.ranks.nl/stopwords/polish

macierzy dokumentów-wyrażeń[7] utworzonej dla korpusu 1, czyli artykułów z zeszytu nr 89. Przedstawia ona liczbę wystąpień poszczególnych słów w danych artykułach np. w artykule 89-37 (tj. artykuł 37 z zeszytu 89) wystąpiło 4 razy słowo „inspekcja" oraz 2 razy słowo „inspektor" a w artykule 89-38 1 raz słowo „inspirowanie".

Następnie dla każdego korpusu obliczono liczbę najczęściej występujących słów oraz ich korelację z innymi słowami. Wyniki zaprezentowano na rysunkach nr 3, 4, 5 i 6.

```
    Terms
Docs auto dom piłka rower ulica
   1    0   0     0     0     0
   2    1   0     0     0     0
   3    1   1     0     0     0
   4    1   1     1     0     0
   5    1   1     1     1     0
   6    1   1     1     1     1
> findAssocs(dm, "auto", 0)
$auto
  dom piłka rower ulica
 0.63  0.45  0.32  0.20
```

```
    Terms
Docs auto dom piłka rower ulica
   1    0   0     0     0     0
   2    1   0     0     0     0
   3    1   1     0     0     0
   4    1   3     3     0     0
   5    1   2     3     1     0
   6    1   1     1     1     1
> findAssocs(dm, "auto", 0)
$auto
  dom piłka rower ulica
 0.49  0.39  0.32  0.20
```

Rys. 2. Przykładowe macierz dokumentów-wyrażeń oraz wartość korelacji dla słowa „auto"
Źródło: Opracowanie własne.

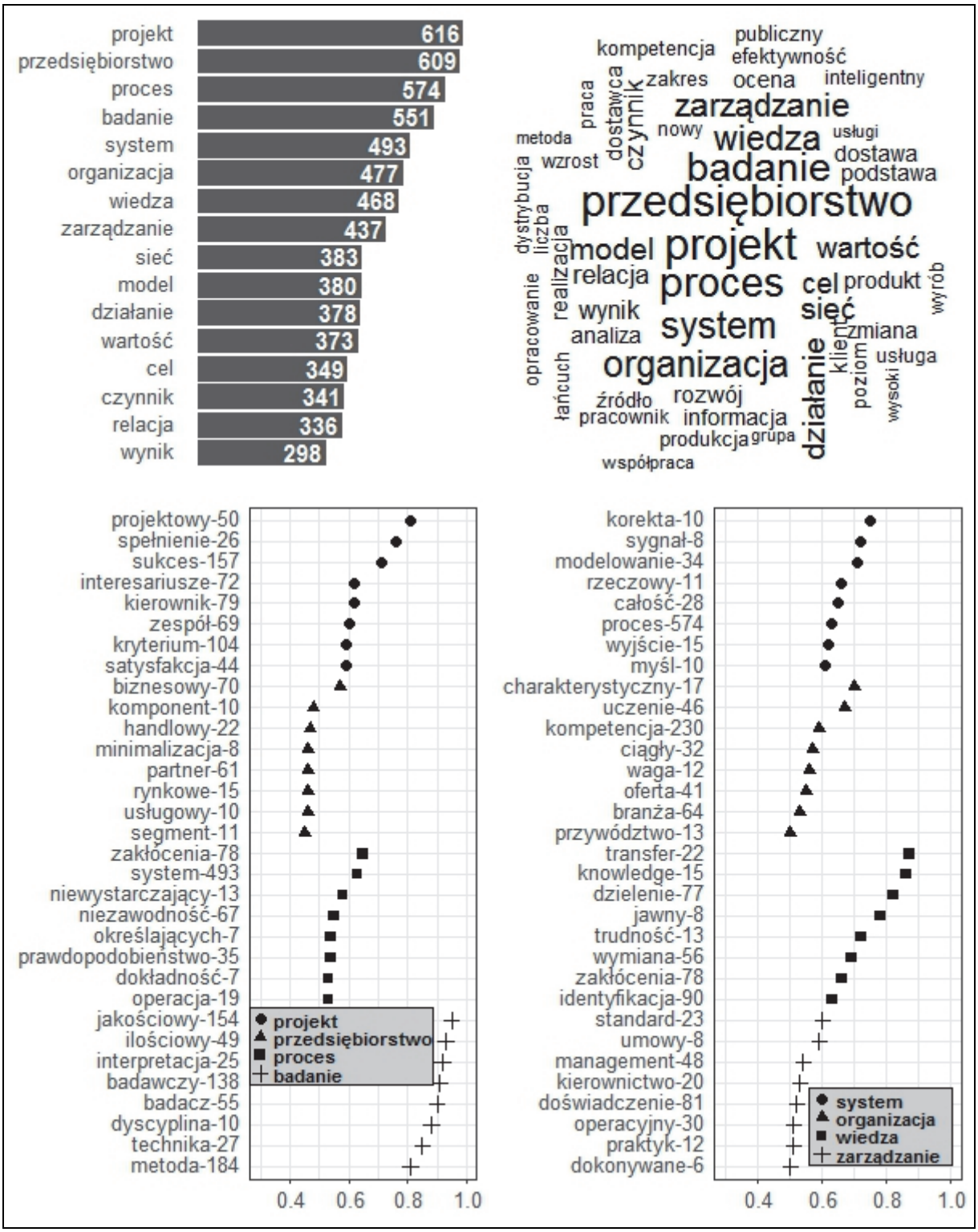
Korelacja pomiędzy słowami została obliczona za pomocą funkcji findAssocs() bazującej na standardowej funkcji cor() z pakietu statystycznego języka R. Na rys. 2 przedstawiono przykładowe dwie macierze dokumentów-wyrażeń oraz obliczoną wartość korelacji dla słowa „auto" z wykorzystanikolem funkcji findAssocs(). W macierzy przedstawionej po lewej stronie rys. 2 słowa wystąpiły tylko raz w danym dokumencie. W macierzy przedstawionej po prawej stronie rys. 2 niektóre ze słów wystąpiły częściej niż jeden raz (np. słowo „dom" wystąpiło 3 razy w dokumencie nr 4, 2 razy w dokumencie nr 5, oraz 1 raz w dokumencie nr 3 i nr 6). W związku z innymi wartościami tych dwóch macierzy wartość korelacji słowa „auto" ze słowami „dom" oraz „piłka" jest inna. Wartość korelacji równa 1 oznacza, że dane dwa słowa zawsze występują razem w dokumentach. Wartość 0 oznacza, że słowa nigdy nie wystąpiły razem.

Prawa górna część rysunków nr 3, 4, 5 oraz 6 przedstawia w postaci tzw. chmury słów[8], pięćdziesiąt najczęściej używanych słów. Wielkość słów zależy od liczby wystąpień danego słowa. Im dane słowo częściej występowało w analizowanym korpusie tym jest większe. Doskonale widać to na chmurze słów wykonanej dla zeszytu nr 95 (rys.4), gdzie górują słowa „społeczny" oraz „rozwój" – wstąpiły one odpowiednio 768 i 687 razy.

W dolnej części rysunków nr 3, 4, 5 oraz 6 przedstawiono dla ośmiu najczęściej używanych słów ich związek (na bazie obliczonego współczynnika korelacji) z innymi słowami użytymi w danym korpusie. Lewa strona przedstawia korelację dla pierwszych

[7] Oryginalna macierz ma rozmiary 42 x 10315 (42-dokumenty, 10315 - wyrażeń (słów)).
[8] Chmury słów były tworzone z zastosowaniem tych samych parametrów dla wszystkich korpusów.

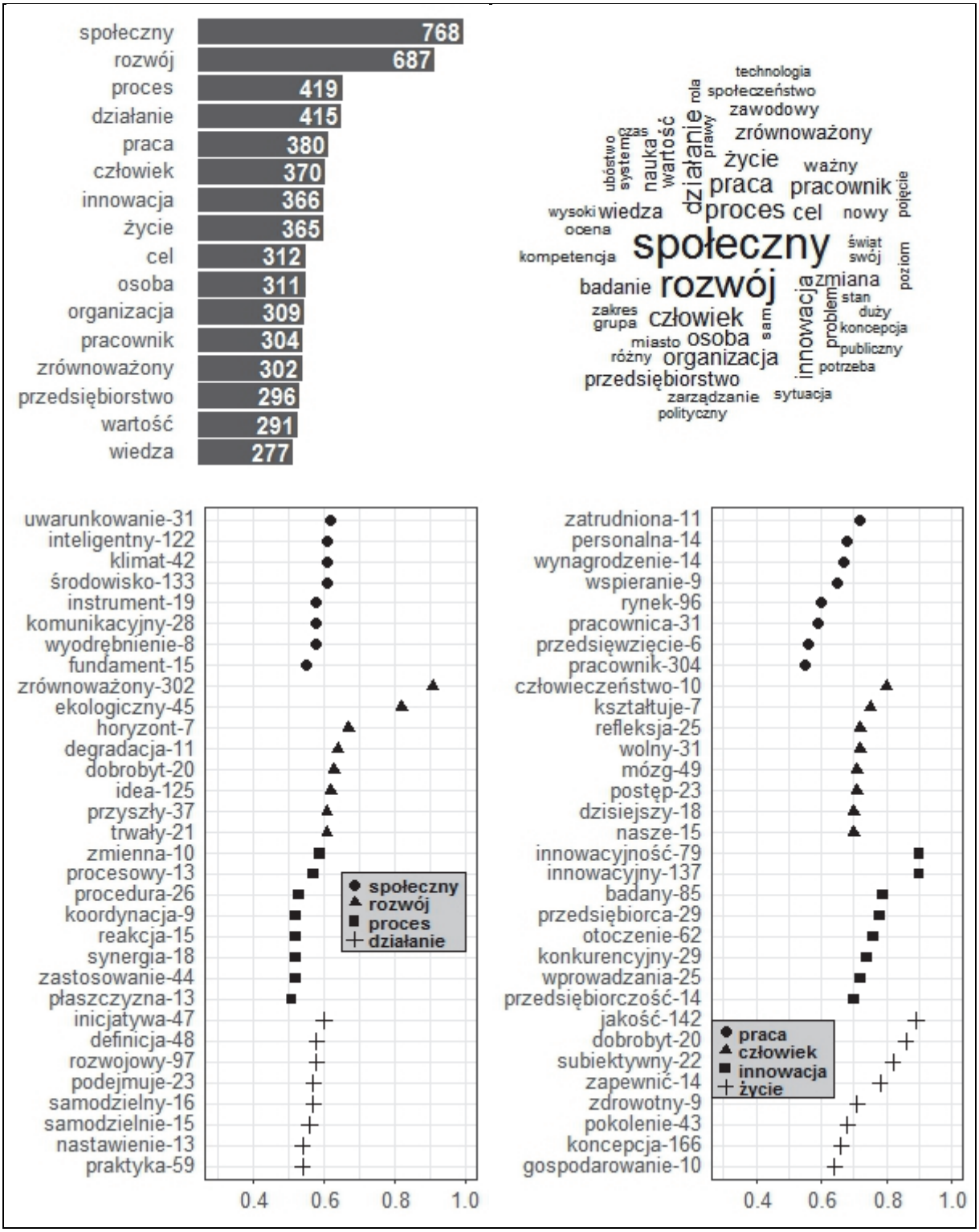
czterech, najczęściej używanych słów a prawa dla kolejnych czterech. Na przykład słowo „badanie" (551 użyć), czyli czwarte najczęściej używane słowo w ZN nr 89 (rys. 3), miało największą korelację ze słowami „jakościowy", „ilościowy", „interpretacja", „badawczy" itd.[9]



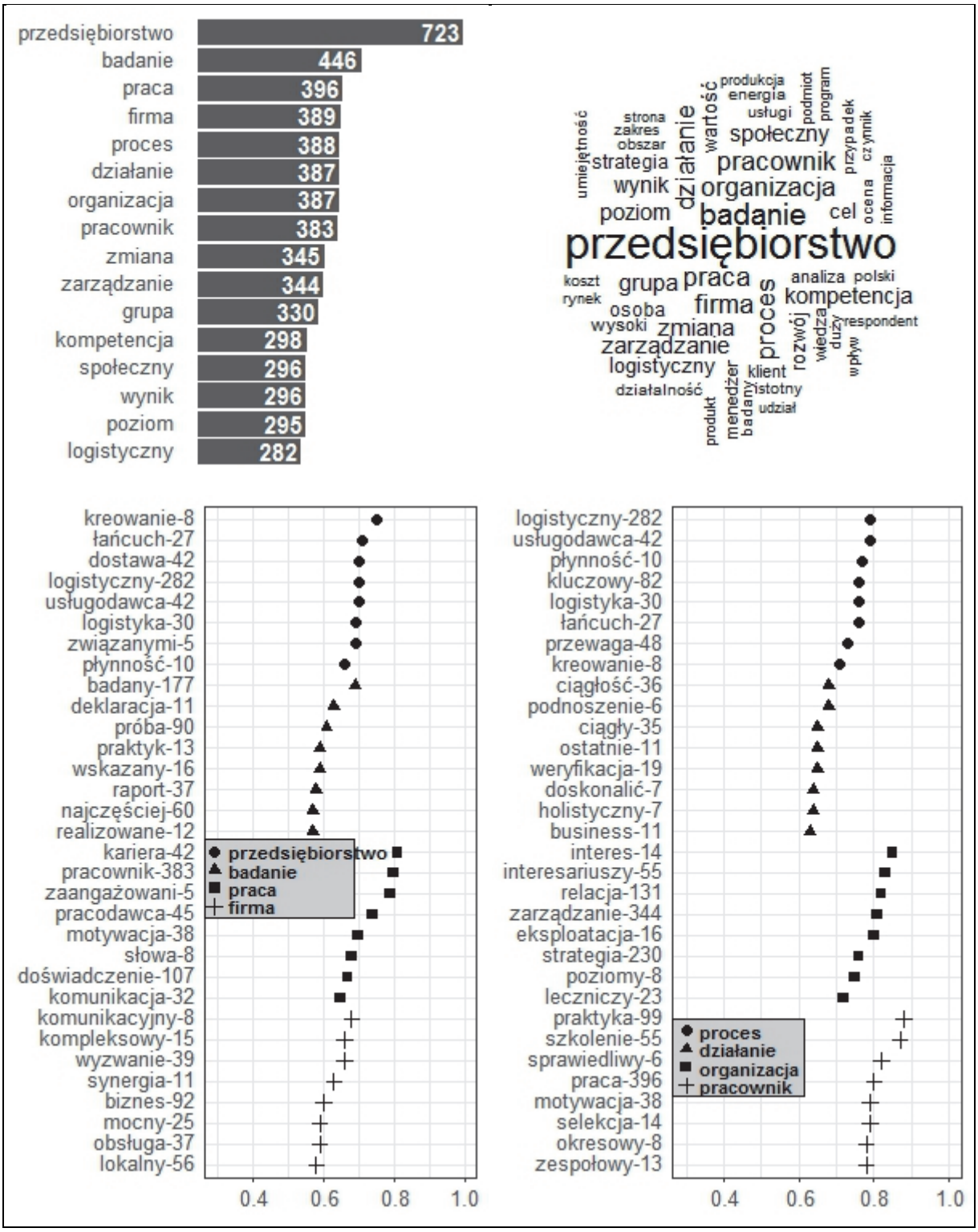
Rys. 3. Analiza korpusu nr 1
Źródło: Opracowanie własne.

[9] znajdujące się obok tych słów liczby informują ile razy dane słowo zostało użyte w korpusie np. słowo „jakościowy" - 154 a „ilościowy" - 49 razy.

Wartość korelacji dla słowa „badanie" przedstawia znak „+", słowa „projekt" znak „●", słowa „przedsiębiorstwo" znak „▲" a słowa „proces" znak „■". Dla lepszej przejrzystości, dane na wykresie posortowano w porządku malejącym według liczby wystąpień danego słowa w korpusie, a następnie według wartości jego korelacji z innymi słowami.



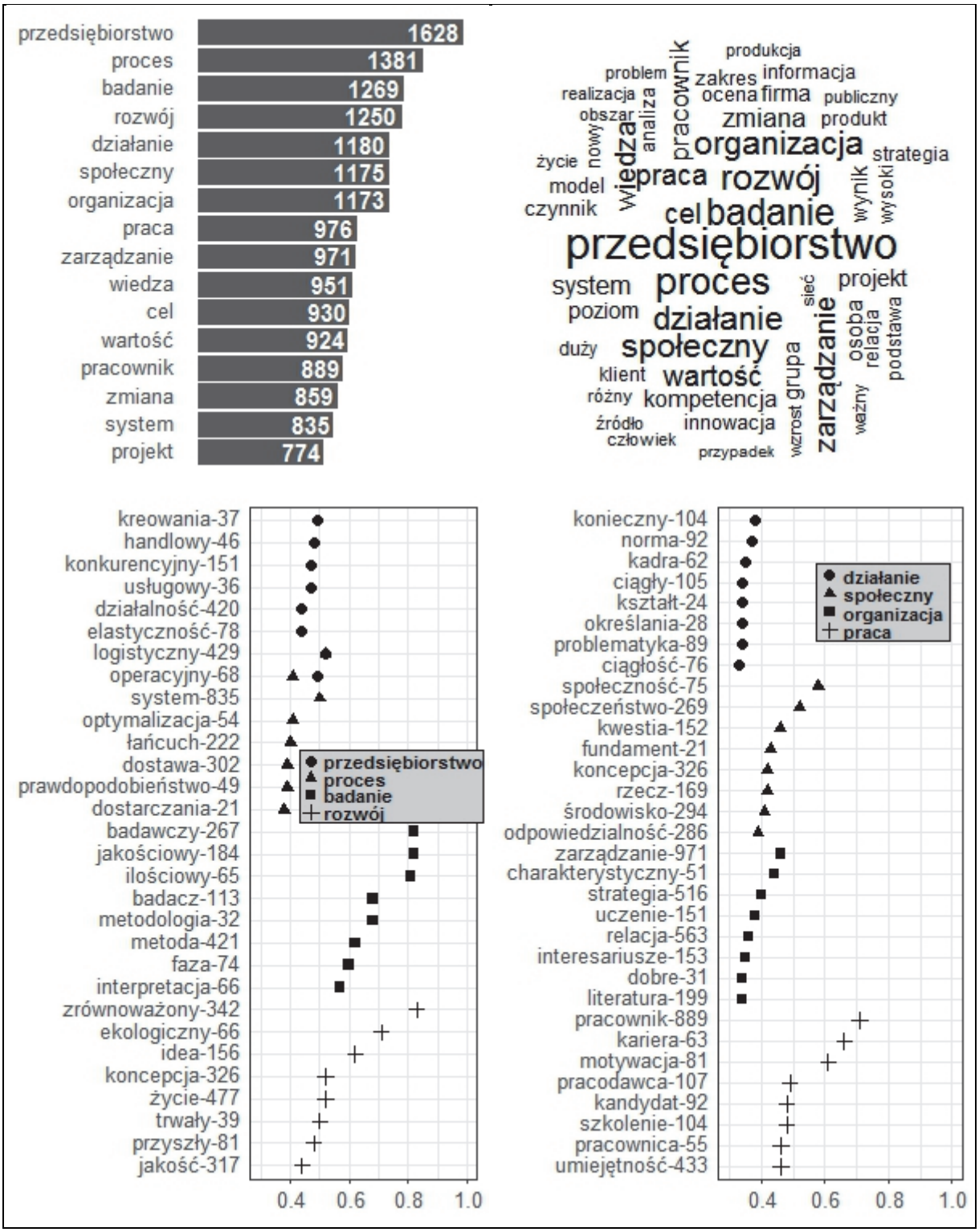
Rys. 4. Analiza korpusu nr 2
Źródło: Opracowanie własne.



Rys. 5. Analiza korpusu nr 3
Źródło: Opracowanie własne.

Według wyników przedstawionych na rys. 6 najczęściej używanym, przez autorów słowem w korpusie nr 4 było „przedsiębiorstwo". Słowo to było również najczęściej używanym słowem w korpusie nr 3 (rys. 5). W korpusie nr 1 było na miejscu drugim (rys. 3) a w korpusie nr 2 na miejscu czternastym (rys. 4). Drugie najczęściej używane słowo

z korpusu nr 4 to słowo „proces". Było ono na miejscu trzecim w korpusie nr 1 (rys. 3) i nr 2 (rys. 4) oraz na piątym w korpusie nr 3 (rys. 5).



Rys. 6. Analiza korpusu nr 4
Źródło: Opracowanie własne.

Warto również przyjrzeć się korelacji najczęściej używanych słów. Na przykład gdy w artykule wystąpiło słowo „projekt" (rys. 3) to używane były również m.in. takie słowa jak „projektowy", „sukces", „interesariusze", „kierownik", „zespół", gdy wystąpiło słowo

„przedsiębiorstwo” (rys. 3) to towarzyszyły mu m.in. słowa „biznesowy”, „handlowy”, „minimalizacja”, „partner”, „usługowy” a ze słowem „rozwój” (rys. 4) występowały m.in. słowa „zrównoważony” i „ekologiczny”.

Na rys. 6 można zauważyć że dwa najczęściej używane słowa z korpusu nr 4 czyli „przedsiębiorstwo” i „proces”, są równocześnie skorelowane ze słowami „logistyczny” (ikony nakładają się na siebie) i „operacyjny” (dwie ikony w jednej linii) - jedyny taki przypadek na prezentowanych wykresach.

Autorami części artykułów analizowanych w ramach korpusu nr 2 byli pracownicy Katedry Stosowanych Nauk Społecznych Wydziału Organizacji i Zarządzania Politechniki Śląskiej. Ten fakt a raczej zainteresowania badawcze autorów spowodowały, że najczęściej używanym słowem w tym korpusie było słowo „społeczny”.

W korpusie nr 1 autorami części publikacji byli pracownicy Instytutu Zarządzania, Administracji i Logistyki wspomnianego wydziału. Jak wynika z analizy część autorów zajmuje się najprawdopodobniej tematyką zarządzania projektami, zarządzania wiedzą, zarządzaniem przedsiębiorstwem.

Autorami części artykułów wchodzących w skład korpusu nr 3 byli natomiast pracownicy Instytutu Ekonomii i Informatyki w/w wydziału. Z utworzonej dla tego korpusu chmury słów można wnioskować, że zainteresowania badawcze autorów są generalnie skupione na badaniu przedsiębiorstw.

Oczywiście w skład każdego z analizowanych korpusów wchodziły także artykuły osób prowadzących badania w ramach zewnętrznych jednostek naukowych. Ponieważ część zeszytów naukowych jest „efektem ubocznym” organizowanych konferencji i seminariów naukowych to można przyjąć, że zainteresowania badawcze autorów z zewnętrznych jednostek naukowych i autorów z Wydziału Organizacji i Zarządzania w ramach danego korpusu dotyczyły wspólnych obszarów tematycznych.

4. Zakończenie

Przeprowadzona analiza text mining pozwoliła pozyskać nowe wcześniej nie znane informacje na temat analizowanych publikacji. Uzyskano informacje na temat liczby najczęściej używanych słów oraz ich korelacji z innymi słowami, które przedstawiono w formie graficznej.

Przeprowadzając tego typu analizę trzeba podjąć kilka istotnych decyzji. Jedną z nich jest np. pytanie: Czy usunąć ewentualne synonimy? Czy słowo „przedsiębiorstwo” powinno zostać scalone w jedno słowo, ze słowem „organizacja”? Słowo „organizacja” (rys. 6) mogło przecież zostać użyte w różnych kontekstach. Mogło występować jako rzeczownik np. „organizacja gospodarcza” (i w tym przypadku mogłoby zostać potraktowane jako synonimi

słowa „przedsiębiorstwo") ale również jak czasownik w wyrażeniu „organizacja zespołu projektowego". Istotnym zagadnieniem jest także stworzenie listy tzw. stop-words, w celu usunięcia nieprzydatnych w analizie słów – w przedstawionej analizie usunięto np. słowo „źródło" użyte w celu wskazania pozycji literaturowej .

W przypadku analizy tekstu w języku polskim należy również rozważyć co zrobić ze znajdującymi się w treści słowami w języku obcym np. „business intelligence" – usunąć, zostawić, a może przetłumaczyć na język polski.

Bibliografia

1. Berry, M.W., Kogan, J.: Text mining: applications and theory. John Wiley & Sons. 2010.
2. Gładysz A., Zastosowanie metod eksploracyjnej analizy tekstu w logistyce., Logistyka No 3, 2012, s. 643–651.
3. Hearst M.A.: Untangling Text Data Mining, Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999. Dostępny 17.05.2017: http://dl.acm.org/citation.cfm?id=1034679
4. Kao A., Poteet S.: Natural Language Processing and Text Mining, London, Springer 2007.
5. Lula P., Text mining jako narzędzie pozyskiwania informacji dokumentów tekstowych. 2015. Dostępny 17.05.2017: http://media.statsoft.nazwa.pl/_old_dnn/downloads/text_mining_jako_narzedzie_pozyskiwania.pdf.
6. Mirończuk M.: Przegląd metod i technik eksploracji danych tekstowych." Studia i Materiały Informatyki Stosowanej, Tom 4, Nr 6, 2012.
7. Tan A.: Text Mining: The state of the art and the challenges, Pacific Asia Conference on Knowledge Discovery and Data Mining PAKDA. 1999.