PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 327

Taksonomia 22
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga, Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Beata Mazur
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp 9

Eugeniusz Gatnar, Balance of payments statistics and external competitiveness of Poland 15

Andrzej Sokołowski, Magdalena Czaja, Efektywność metody *k*-średnich w zależności od separowalności grup 23

Barbara Pawełek, Józef Pociecha, Adam Sagan, Wielosektorowa analiza ukrytych przejść w modelowaniu zagrożenia upadłością polskich przedsiębiorstw 30

Elżbieta Gołata, Zróżnicowanie procesu starzenia i struktur demograficznych w Poznaniu i aglomeracji poznańskiej na tle wybranych dużych miast Polski w latach 2002-2011 39

Aleksandra Łuczak, Feliks Wysocki, Ustalanie systemu wag dla cech w zagadnieniach porządkowania liniowego obiektów 49

Marek Walesiak, Wzmacnianie skali pomiaru dla danych porządkowych w statystycznej analizie wielowymiarowej 60

Paweł Lula, Identyfikacja słów i fraz kluczowych w tekstach polskojęzycznych za pomocą algorytmu *RAKE* 69

Mariusz Kubus, Propozycja modyfikacji metody złagodzonego LASSO 77

Andrzej Bąk, Tomasz Bartłomowicz, Wielomianowe modele logitowe wyborów dyskretnych i ich implementacja w pakiecie `DiscreteChoice` programu R 85

Justyna Brzezińska, Wykorzystanie modeli logarytmiczno-liniowych do analizy bezrobocia w Polsce w latach 2004-2012 95

Andrzej Bąk, Marcin Pełka, Aneta Rybicka, Zastosowanie pakietu `dcMNM` programu R w badaniach preferencji konsumentów wódki 104

Barbara Batóg, Jacek Batóg, Analiza stabilności klasyfikacji polskich województw według sektorowej wydajności pracy w latach 2002-2010 113

Małgorzata Markowska, Danuta Strahl, Klasyfikacja europejskiej przestrzeni regionalnej ze względu na filary inteligentnego rozwoju z wykorzystaniem referencyjnego systemu granicznego 121

Kamila Migdał-Najman, Krzysztof Najman, Formalna ocena jakości odwzorowania struktury grupowej na mapie Kohonena 131

Kamila Migdał-Najman, Krzysztof Najman, Graficzna ocena jakości odwzorowania struktury grupowej na mapie Kohonena 139

Beata Basiura, Anna Czapkiewicz, Badanie jakości klasyfikacji szeregów czasowych 148

Michał Trzęsiok, Wybrane metody identyfikacji obserwacji oddalonych 157

Grażyna Dehnel, Tomasz Klimanek, Taksonomiczne aspekty estymacji pośredniej uwzględniającej autokorelację przestrzenną w statystyce gospodarczej ... 167
Michał Bernard Pietrzak, Justyna Wilk, Odległość ekonomiczna w modelowaniu zjawisk przestrzennych z wykorzystaniem modelu grawitacji ... 177
Maciej Beręsewicz, Próba zastosowania różnych miar odległości w uogólnionym estymatorze Petersena ... 186
Marcin Szymkowiak, Tomasz Józefowski, Konstrukcja i praktyczne wykorzystanie estymatorów typu SPREE na przykładzie dwuwymiarowych tabel kontyngencji ... 195
Marcin Pełka, Klasyfikacja pojęciowa danych symbolicznych w podejściu wielomodelowym ... 202
Małgorzata Machowska-Szewczyk, Ocena klas w rozmytej klasyfikacji obiektów symbolicznych ... 210
Justyna Wilk, Problem wyboru liczby klas w taksonomicznej analizie danych symbolicznych ... 220
Andrzej Dudek, Metody analizy skupień w klasyfikacji markerów map Google ... 229
Ewa Roszkowska, Ocena ofert negocjacyjnych w słabo ustrukturyzowanych problemach negocjacyjnych z wykorzystaniem rozmytej procedury SAW 237
Marcin Szymkowiak, Marek Witkowski, Zastosowanie analizy korespondencji do badania kondycji finansowej banków spółdzielczych ... 248
Bartłomiej Jefmański, Budowa rozmytych indeksów satysfakcji klientów z zastosowaniem programu R ... 257
Karolina Bartos, Odkrywanie wzorców zachowań konsumentów za pomocą analizy koszykowej danych transakcyjnych ... 266
Joanna Trzęsiok, Taksonomiczna analiza krajów pod względem dzietności kobiet oraz innych czynników demograficznych ... 275
Beata Bal-Domańska, Próba identyfikacji większych skupisk regionalnych oraz ich konwergencja ... 285
Beata Bieszk-Stolorz, Iwona Markowicz, Wpływ zasiłku na proces poszukiwania pracy ... 294
Marta Dziechciarz-Duda, Klaudia Przybysz, Wykształcenie a potrzeby rynku pracy. Klasyfikacja absolwentów wyższych uczelni ... 303
Tomasz Klimanek, Problem pomiaru procesu dezagraryzacji wsi polskiej w świetle wielowymiarowych metod statystycznych ... 313
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska, Wybrane metody analizy danych wzdłużnych ... 321
Artur Zaborski, Zastosowanie miar odległości dla danych porządkowych do agregacji preferencji indywidualnych ... 330
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek, Zastosowanie analizy korespondencji do badania sytuacji mieszkańców strefy podmiejskiej Warszawy na rynku pracy ... 338

Katarzyna Wawrzyniak, Klasyfikacja województw według stopnia realizacji priorytetów Strategii Rozwoju Kraju 2007-2015 z wykorzystaniem wartości centrum wierszowego 346

Summaries

Eugeniusz Gatnar, Statystyka bilansu płatniczego a konkurencyjność gospodarki Polski 22
Andrzej Sokołowski, Magdalena Czaja, Cluster separability and the effectiveness of *k*-means method 29
Barbara Pawełek, Józef Pociecha, Adam Sagan, Multisectoral analysis of latent transitions in bankruptcy prediction models..... 38
Elżbieta Gołata, Differences in the process of aging and demographic structures in Poznań and the agglomeration compared to selected Polish cities in the years 2002-2011 48
Aleksandra Łuczak, Feliks Wysocki, Determination of weights for features in problems of linear ordering of objects 59
Marek Walesiak, Reinforcing measurement scale for ordinal data in multivariate statistical analysis 68
Paweł Lula, Automatic identification of keywords and keyphrases in documents written in Polish..... 76
Mariusz Kubus, The proposition of modification of the relaxed LASSO method..... 84
Andrzej Bąk, Tomasz Bartłomowicz, Microeconomic multinomial logit models and their implementation in the `DiscreteChoice` R package. 94
Justyna Brzezińska, The analysis of unemployment data in Poland in 2004--2012 with application of log-linear models 103
Andrzej Bąk, Marcin Pełka, Aneta Rybicka, Application of the MMLM package of R software for vodka consumers preference analysis..... 112
Barbara Batóg, Jacek Batóg, Analysis of the stability of classification of Polish voivodeships in 2002-2010 according to the sectoral labour productivity 120
Małgorzata Markowska, Danuta Strahl, Classification of the European regional space in terms of smart growth pillars using the reference limit system..... 130
Kamila Migdał Najman, Krzysztof Najman, Formal quality assessment of group structure mapping on the Kohonen's map 138
Kamila Migdał Najman, Krzysztof Najman, Graphical quality assessment of group structure mapping on the Kohonen's map 147
Beata Basiura, Anna Czapkiewicz, Validation of time series clustering 156
Michał Trzęsiok, Selected methods for outlier detection..... 166

Grażyna Dehnel, Tomasz Klimanek, Taxonomic aspects of indirect estimation accounting for spatial correlation in enterprise statistics 176
Michał Bernard Pietrzak, Justyna Wilk, Economic distance in modeling spatial phenomena with the application of gravity model..... 185
Maciej Beręsewicz, An attempt to use different distance measures in the Generalized Petersen estimator 194
Marcin Szymkowiak, Tomasz Józefowski, Construction and practical using of SPREE estimators for two-dimensional contingency tables..... 201
Marcin Pełka, The ensemble conceptual clustering for symbolic data..... 209
Małgorzata Machowska-Szewczyk, Evaluation of clusters obtained by fuzzy classification methods for symbolic objects..... 219
Justyna Wilk, Problem of determining the number of clusters in taxonomic analysis of symbolic data 228
Andrzej Dudek, Clustering techniques for Google maps markers..... 236
Ewa Roszkowska, The evaluation of negotiation offers in ill structure negotiation problems with the application of fuzzy SAW procedure 247
Marcin Szymkowiak, Marek Witkowski, The use of correspondence analysis in analysing the financial situation of cooperative banks..... 256
Bartłomiej Jefmański, The construction of fuzzy customer satisfaction indexes using R program..... 265
Karolina Bartos, Discovering patterns of consumer behaviour by market basket analysis of the transactional data..... 274
Joanna Trzęsiok, Cluster analysis of countries with respect to fertility rate and other demographic factors 284
Beata Bal-Domańska, An attempt to identify major regional clusters and their convergence 293
Beata Bieszk-Stolorz, Iwona Markowicz, The influence of benefit on the job finding process 302
Marta Dziechciarz-Duda, Klaudia Przybysz, Education and labor market needs. Classification of university graduates 312
Tomasz Klimanek, The problem of measuring deagrarianisation process in rural areas in Poland using multivariate statistical methods..... 320
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska, Selected methods for an analysis of longitudinal data..... 329
Artur Zaborski, The application of distance measures for ordinal data for aggregation individual preferences 337
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek, Application of correspondence analysis to examine the situation of the inhabitants of Warsaw suburban area in the labour market 345
Katarzyna Wawrzyniak, Classification of voivodeships according to the level of the realization of priorities of *the National Development Strategy 2007-2015* with using the values of centroid of the rows 355

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 327 • 2014
Taksonomia 22 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Paweł Lula
Uniwersytet Ekonomiczny w Krakowie

IDENTYFIKACJA SŁÓW I FRAZ KLUCZOWYCH W TEKSTACH POLSKOJĘZYCZNYCH ZA POMOCĄ ALGORYTMU *RAKE*

Streszczenie: Tematyka artykułu związana jest z zagadnieniem automatycznej identyfikacji słów oraz fraz kluczowych w dokumentach tekstowych. Po przedstawieniu podstawowych informacji dotyczących stosowanych metod zaprezentowano algorytm *RAKE* (*Rapid Automatic Keyword Extraction*), a następnie zaproponowano sposób jego modyfikacji, mający na celu jego lepsze przystosowanie do specyfiki języka polskiego. Przedstawiono również przykładowe zastosowanie algorytmu.

Słowa kluczowe: identyfikacja słów i fraz kluczowych, textmining, eksploracyjna analiza dokumentów

1. Wstęp

Za słowa kluczowe uznaje się te wyrazy, które w sposób syntetyczny najlepiej charakteryzują zawartość tekstu. Słowa kluczowe można uznać za etykiety pozwalające na właściwe zaklasyfikowanie dokumentu. Zbliżonym pojęciem są frazy kluczowe, które można zdefiniować jako sekwencje słów opisujące zawartość dokumentu. Słowa lub frazy kluczowe mogą być przypisane do dokumentu przez jego autora lub przez inną osobę. W zależności od przyjętego rozwiązania osoba przypisująca słowa lub frazy kluczowe może je dobierać w sposób dowolny lub też powinna dokonać ich wyboru ze ściśle zdefiniowanego zestawu.

Problematyka niniejszego artykułu związana jest z zagadnieniem automatycznego określania słów i fraz kluczowych dla istniejących dokumentów. Zastosowanie tego typu narzędzi pozwala na przypisanie do dokumentów etykiet ułatwiających ich klasyfikację czy wyszukiwanie.

Celem niniejszego opracowania jest zaproponowanie i dokonanie wstępnej oceny metody automatycznej identyfikacji słów i fraz kluczowych w tekstach polskojęzycznych, będącej zmodyfikowaną wersją algorytmu *RAKE*.

Praca składa się z sześciu punktów. W punkcie drugim zawarty jest przegląd i klasyfikacja metod automatycznej identyfikacji słów kluczowych. W kolejnej

części przedstawiono oryginalną wersję metody *RAKE*. Prezentacja jej zmodyfikowanej wersji została zamieszczona w punkcie czwartym. Piąty punkt pracy zawiera prezentację sposobu przetwarzania przykładowego dokumentu. Pracę kończy podsumowanie i spis literatury.

2. Metody automatycznej identyfikacji słów kluczowych

Najprostsze podejście do zagadnienia automatycznej identyfikacji słów i fraz kluczowych określić można mianem metod słownikowych. Polega ono na stworzeniu listy istotnych wyrazów lub fraz i zaproponowaniu algorytmu pozwalającego na ich wyszukiwanie w rozpatrywanym tekście. Zidentyfikowane w ten sposób słowa lub frazy uznawane są za kluczowe.

Pewnym rozwinięciem przedstawionego powyżej podejścia są metody regułowe, w których dopuszczalne jest tworzenie wzorców fraz kluczowych za pomocą dostępnego w danym systemie formalizmu (np. opartego na wyrażeniach regularnych). Definiowane wzorce mogą odwoływać się do wartości znaczników definiujących poszczególne fragmenty tekstu (np. znaczników HTML lub XML), co pozwala większą wagę nadać słowom znajdującym się np. w tytułach rozdziałów lub podrozdziałów czy też w streszczeniu lub podsumowaniu.

Wadą przedstawionych metod jest konieczność ich manualnego definiowania. Cechy tej nie mają statystyczne metody identyfikacji słów i fraz kluczowych, które w zależności od wykorzystywanych sposobów analizy można podzielić na [Gładysz 2013]:

- *metody bazujące na macierzy częstości* – zakłada się w nich zwykle, że kluczowy charakter słów lub fraz związany jest z ich częstszym występowaniem w dokumencie;
- *metody wykorzystujące dekompozycję macierzy częstości według wartości osobliwych* – należy oszacować macierz wariancji-kowariancji dla słów wykorzystując k pierwszych składowych uzyskanych poprzez dekompozycję SVD macierzy częstości; za miarę ważności poszczególnych słów uznaje się elementy diagonalne oszacowanej macierzy wariancji-kowariancji;
- *metody korzystające z wyników analizy taksonomicznej* – za słowa kluczowe uznaje się te, których uwzględnienie pozwala uzyskać optymalny podział dokumentów na skupienia (w sensie przyjętej miary jakości skupień);
- *metody oparte na ukrytej alokacji Dirichleta* – są to wyrazy mające największe znaczenie przy konstruowaniu charakterystyk zidentyfikowanych klas ukrytych;
- *metody bazujące na klasyfikacji wzorcowej* – oparte są na klasyfikatorze budowanym w oparciu o zbiór uczący, zawierający dokumenty i przypisane do nich przez człowieka słowa lub frazy kluczowe;
- *metody grafowe* – pozwalają na reprezentację za pomocą grafu relacji pomiędzy słowami i ich uwzględnienie przy konstruowaniu miar ważności słów;

- *metody wizualizacji* – wykorzystują wyniki metod obliczeniowych do stworzenia graficznej formy prezentacji istotności słów lub fraz (np. w postaci chmury słów).

Klasyfikacja metod identyfikacji słów i fraz kluczowych może uwzględniać również kryteria inne niż stosowane w trakcie ich realizacji metody statystyczne. Do głównych kryteriów podziału można zaliczyć:

1) zakres wykorzystywanej informacji – stosując to kryterium wyróżnić można:

a) metody korzystające wyłącznie z przetwarzanych tekstów, które można podzielić na:

i) korpusowe – jeśli w trakcie przetwarzania dokumentu wykorzystują informacje dotyczące całego przetwarzanego korpusu,

ii) dokumentowe – jeśli zakres informacji wykorzystywanych w trakcie analizy dokumentu ograniczony jest wyłącznie do jego zawartości,

b) metody z bazą wiedzy dziedzinowej, przyjmującej np. postać ontologii;

2) rodzaj generowanych wyników – przyjęcie tego kryterium pozwala na wydzielenie:

a) metod identyfikujących słowa kluczowe,

b) metod identyfikujących frazy kluczowe;

3) typ algorytmu uczącego – pozwala na wyróżnienie:

a) metod działających w trybie z nauczycielem – w tym przypadku konieczne jest zaprezentowanie zbioru dokumentów z przypisanymi manualnie słowami kluczowymi,

b) metod działających w trybie bez nauczyciela – które nie wymagają przykładów zawierających przypisane słowa lub frazy kluczowe.

3. Opis algorytmu *RAKE*

Algorytm *RAKE* (*Rapid Automatic Keyword Extraction*) został zaproponowany w pracy [Rose i in. 2010]. Pozwala na identyfikację słów i fraz kluczowych w dokumentach tekstowych. Nie korzysta z wiedzy dziedzinowej. W trakcie identyfikacji operuje wyłącznie na pojedynczym dokumencie i nie uwzględnia zawartości innych dokumentów wchodzących w skład przetwarzanego korpusu. Działa w trybie bez nauczyciela, czyli nie wymaga podawania przykładów prawidłowo zdefiniowanych słów kluczowych. Jego działanie składa się z kilku etapów.

Etap 1. Podział tekstu na frazy kandydujące

Przez frazy kandydujące rozumie się wydzielone w tekście słowa lub sekwencje słów będące potencjalnymi słowami lub frazami kluczowymi. Podział tekstu przeprowadzany jest w dwóch krokach:

- podział w miejscu wystąpienia separatorów (kropka, średnik, przecinek, myślnik, pytajnik, wykrzyknik),
- podział sekwencji słów uzyskanych w powyższym kroku w miejscu wystąpienia słów nieistotnych (będących elementami stop-listy, np. ale, to, niech, się,

gdyż, na itd.). Zawartość stop-listy jest jednym z ważniejszych parametrów sterujących działaniem algorytmu.

Etap 2. Budowa grafu współwystępowania

Graf współwystępowania reprezentuje związki pomiędzy wyrazami dokumentu, wynikające z faktu wspólnego występowania we frazach kandydujących. Węzły grafu reprezentują słowa występujące we frazach kandydujących. W węźle reprezentującym *i*-te słowo przechowywana jest informacja dotycząca liczby jego wystąpień we frazach kandydujących (wartość $freq(w_i)$). Istnienie krawędzi pomiędzy węzłami odpowiadającymi *i*-temu oraz *j*-temu słowu wskazuje, że słowa te pojawiły się łącznie przynajmniej raz w tej samej frazie kandydującej. Waga przypisana do krawędzi określa dokładną liczbę współwystąpień we frazach kandydujących rozpatrywanej pary słów. Wartość stopnia węzła odpowiadającego *i*-temu słowu (czyli wartość $deg(w_i)$) wskazuje łączną liczbę współwystąpień danego słowa z innymi słowami.

Etap 3. Ocena słów i fraz kandydujących

Ocena rozpoczyna się od wyznaczenia miary ważności każdego słowa występującego we frazach kandydujących. Autorzy algorytmu proponują następujące miary:

$$score(w_i) = freq(w_i),$$

$$score(w_i) = freq(w_i) + deg(w_i),$$

$$score(w_i) = \big(freq(w_i) + deg(w_i)\big)/freq(w_i).$$

Po wyznaczeniu wartości słów obliczana jest ocena fraz. Jest ona równa sumie ocen przypisanych do wyrazów wchodzących w skład frazy.

Etap 4. Identyfikacja fraz złożonych, zawierających słowa uznane za nieistotne

Przedstawiony w opisie pierwszego etapu przetwarzania sposób podziału tekstu na frazy kandydujące zakładał, że miejsce podziału jest wyznaczane przez słowa nieistotne, które nie wchodziły w skład wyodrębnionych fragmentów. Z tego powodu słowa wchodzące w skład stop-listy nigdy nie mogłyby pojawić się jako element fraz kandydujących. W celu rozwiązania tej niedogodności algorytm *RAKE* przewiduje, że jeśli sekwencja złożona z przynajmniej dwóch fraz kandydujących i rozdzielających je słów nieistotnych wystąpi w dokumencie przynajmniej dwukrotnie, to tworzy ona nową frazę kandydującą (wraz z zawartymi w niej elementami stop-listy).

Ocena utworzonej w ten sposób nowej frazy jest wyznaczana poprzez zsumowanie wartości ocen wchodzących w jej skład słów.

Etap 5. Sporządzenie listy fraz wynikowych

Wynikiem działania algorytmu *RAKE* jest lista fraz kandydujących uporządkowanych malejąco według wyznaczonych ocen. Autorzy metody proponują, aby

uwzględnić w charakterze fraz kluczowych początkowe elementy tak utworzonej listy w liczbie równej jednej trzeciej wszystkich słów występujących w analizowanym dokumencie.

Najbardziej znaną implementacją algorytmu RAKE jest program napisany w języku Python i dostępny w ramach projektu: https://github.com/aneesha/RAKE. Algorytm jest również oprogramowany przy wykorzystaniu pakietu *Natural Language Toolkit* (http://www.nltk.org) [Perkins 2010], zaś kod dostępny jest pod adresem: http://sujitpal.blogspot.com/2013/03/implementing-rake-algorithm-with-nltk.html.

4. Propozycja modyfikacji algorytmu *RAKE*

Jedną z zalet algorytmu jest jego uniwersalność przejawiająca się możliwością przetwarzania dokumentów napisanych w dowolnym języku. Postulat ten jest spełniony tylko częściowo w przypadku języków fleksyjnych, w których wielość możliwych form wyrazów utrudnia prawidłowe działanie metody. Tego typu problemy są szczególnie widoczne w przypadku próby zastosowania algorytmu *RAKE* do analizy tekstów polskojęzycznych. Stwarzają one konieczność modyfikacji algorytmu w celu dostosowania jego działania do specyfiki języka polskiego.

Opracowując zmodyfikowaną wersję algorytmu *RAKE*, wprowadzono kilka zmian w stosunku do wersji oryginalnej. Ich listę przedstawia niniejsza część pracy.

A. Zastosowanie polskiej stop-listy

Jest to oczywista zmiana wynikająca z potrzeby przystosowania algorytmu do pracy na dokumentach polskojęzycznych. Na potrzeby obliczeń przygotowano ogólną, stosunkowo niewielką, listę słów nieistotnych.

B. Zmodyfikowany sposób oceny słów

Proces wyznaczania ocen poszczególnych wyrazów poprzedzony został przekształceniem poszczególnych wyrazów do ich formy podstawowej[1]. Forma podstawowa wykorzystywana była w trakcie realizacji obliczeń, natomiast w trakcie wyświetlania wyników wyrazy i frazy prezentowane były w formie występującej w dokumencie.

Dla każdego wyrazu wyznaczone zostały trzy oceny cząstkowe:

a) liczba wystąpień wyrazu w dokumencie ($freq(w_i)$),

b) liczba różnych wyrazów, z którymi rozpatrywany wyraz występuje w tych samych frazach ($degree1(w_i)$),

c) liczba różnych wyrazów, z którymi rozpatrywany wyraz występuje w tych samych fraz więcej niż jeden raz ($degree2(w_i)$).

[1] W celu przekształcenia do formy podstawowej wyrazów w tekstach polskojęzycznych zastosowano bibliotekę *Morfologik* (http://morfologik.blogspot.com/).

Uzyskane oceny cząstkowe zostały przeskalowane do przedziału [0; 1] poprzez podzielenie uzyskanych wartości przez wyznaczone wartości maksymalne dla *freq*, *degree1* oraz *degree2*. Następnie zostały one zsumowane dla poszczególnych wyrazów. Uzyskane w ten sposób końcowe oceny wyrazów przyjmowały wartości rzeczywiste z przedziału [0; 3].

C. Zmodyfikowany sposób tworzenia fraz kandydujących

Po dokonaniu podziału tekstu na fragmenty w miejscach pojawienia się separatorów lub słów nieistotnych uzyskano sekwencje słów, które stanowiły podstawę do zdefiniowania fraz kandydujących. Za frazę kandydującą uznano każdy fragment sekwencji umieszczonej pomiędzy separatorami.

Na przykład, jeśli pomiędzy separatorami znalazł się tekst: *dynamiczna analiza danych wielowymiarowych*, to w charakterze fraz kandydujących przyjęte zostały sekwencje: *dynamiczna*; *dynamiczna analiza*; *dynamiczna analiza danych*; *dynamiczna analiza danych wielowymiarowych*; *analiza*; *analiza danych*; *analiza danych wielowymiarowych*; *danych*; *danych wielowymiarowych*; *wielowymiarowych*

Warto podkreślić, że utworzona w poprzednim kroku stop-lista spełnia odmienną rolę niż w oryginalnej wersji algorytmu. Wyrazy ze stop-listy nigdy nie pojawią się we frazach kluczowych. Z tego powodu dobór jej elementów musi być przemyślany, a jej zawartość niewielka. Należy też pamiętać, że ograniczniki fraz kandydujących mogą się pojawić w dowolnym miejscu w tekście, a nie tylko w miejscu separatorów.

D. Zmodyfikowany sposób tworzenia listy wynikowej

Wstępnie utworzona lista fraz kandydujących uporządkowanych malejąco według wyznaczonych ocen poddana została dalszemu przekształceniu. Polegało ono na wyznaczeniu dla każdej frazy zbioru zawartych w niej słów przekształconych do postaci podstawowej i usunięciu tych fraz, których tak zdefiniowany zbiór stanowił podzbiór zbioru przypisanego jednej z fraz znajdujących się na wyższych pozycjach listy. Na przykład, jeśli wstępna lista fraz zawierała trzy elementy:

analiza zachowań konsumentów (analiza, *zachowanie*, *konsument)*,

model zachowania konsumenta (model, *zachowanie*, *konsument)*,

modele zachowań (model, *zachowanie)*,

to trzeci jej element zostanie usunięty.

5. Przykładowe zastosowanie metody

W celu zaprezentowania sposobu funkcjonowania algorytmu dokonano identyfikacji słów i fraz kluczowych w tekście będącym streszczeniem referatu przygotowanego z myślą o konferencji SKAD 2013. Streszczenie to miało następującą postać: *Wzmacnianie skali pomiaru dla danych porządkowych. W statystycznej analizie*

wielowymiarowej punktem wyjścia zastosowania metod statystycznej analizy wielowymiarowej jest macierz danych. Problem stosowania narzędzi statystycznej analizy wielowymiarowej komplikuje się wtedy, gdy w zbiorze znajdują się zmienne mierzone na skalach różnych rodzajów lub tylko na słabych skalach pomiaru (szczególnie na skali porządkowej). W artykule proponuje się metodę wzmacniania skali pomiaru zmiennych porządkowych. Propozycja bazuje na odległości gdm2 właściwej do zastosowania dla danych porządkowych. Rozważane zagadnienia zilustrowano przykładem empirycznym z wykorzystaniem programu R[2].

W pierwszym etapie obliczeń wyznaczono oceny dla poszczególnych słów. Najwyżej ocenione zostały wyrazy:

analizie: 2,76; *analizy*: 2,76; *statystycznej*: 2,72; *wielowymiarowej*: 2,58; *metodę*: 2,20; *metod*: 2,20; *zastosowania*: 2,05; *skali*: 1,94; *skalach*: 1,94; *wyjścia*: 1,73; *punktem*: 1,64; *pomiaru*: 1,43; *porządkowych*: 1,07; *danych*: 1,03; *wzmacnianie*: 1,02; *wzmacniania*: 1,02.

Natomiast końcowa lista wyników uporządkowana malejąco według wartości ocen fraz zawierała elementy:

statystycznej analizy wielowymiarowej: 5,38; *metod statystycznej analizie*: 5,13; *analizie wielowymiarowej punktem*: 4,66; *zastosowania metod statystycznej*: 4,65; *narzędzi statystycznej analizy*: 4,30; *analizy wielowymiarowej komplikuje*: 4,02; *wyjścia zastosowania metod*: 3,99; *wielowymiarowej punktem wyjścia*: 3,97; *metod statystycznej analizie wielowymiarowej*: 3,85; *zastosowania metod statystycznej analizie*: 3,65; *statystycznej analizie wielowymiarowej punktem*: 3,64; *punktem wyjścia zastosowania*: 3,62; *metodę wzmacniania skali*: 3,45.

6. Podsumowanie

Algorytm *RAKE* należy do grupy najlepszych metod identyfikacji fraz kluczowych. Jednakże charakter języka polskiego stwarza konieczność wprowadzenia modyfikacji polegającej na uwzględnieniu polskiej stop-listy i u wzięcia pod uwagę form podstawowych wyrazów. Zmodyfikowano również sposób tworzenia fraz kandydujących i listy wyników końcowych. Analiza sposobu funkcjonowania algorytmu i przeprowadzone eksperymenty wskazują na jeszcze jedną zaletę algorytmu *RAKE*. Jest nią elastyczność metody i łatwość modyfikowania jej kolejnych etapów w celu dostosowania do specyfiki używanego języka i charakteru tekstu.

Wstępna ocena możliwości algorytmu wskazuje na jego przydatność w zastosowaniach praktycznych.

[2] Jest to streszczenie referatu przygotowanego na konferencję SKAD 2013: M. Walesiak, *Wzmacnianie skali pomiaru dla danych porządkowych w statystycznej analizie wielowymiarowej*, Uniwersytet Ekonomiczny we Wrocławiu. Tekst został zaprezentowany w postaci pozbawionej znaków formatujących, przygotowanej do przetworzenia przez program będący implementacją algorytmu *RAKE*.

Literatura

Gładysz A. (2013), *Badanie skuteczności metod identyfikacji słów kluczowych w polskojęzycznych tekstach*, rozprawa doktorska, Uniwersytet Ekonomiczny w Krakowie, Kraków.

Konferencja (2013), XXII Konferencja Naukowa Sekcji Klasyfikacji i Analizy Danych PTS oraz XXVII Konferencja Taksonomiczna nt. „Klasyfikacja i analiza danych – teoria i zastosowania", *Program i streszczenia*, red. M. Walesiak, Uniwersytet Ekonomiczny we Wrocławiu, Wrocław.

Perkins J. (2010), *Python Text Processing with NLTK 2.0 Cookbook*, Packt Publishing.

Rose S., Engel D., Cramer N., Cowley W. (2010), *Automatic Keyword Extraction from Individual Documents*, [w:] M.W. Berry, J. Kogan (red.), *Text Mining: Theory and Applications*, John Wiley & Sons, s. 3-19.

Źródła internetowe

https://github.com/aneesha/RAKE.

http://sujitpal.blogspot.com/2013/03/implementing-rake-algorithm-with-nltk.html.

AUTOMATIC IDENTIFICATION OF KEYWORDS AND KEYPHRASES IN DOCUMENTS WRITTEN IN POLISH

Summary: In the paper the problem of automatic identification of keywords and keyphrases in text documents written in Polish is presented. First, the classification of different approaches to the problem of keywords extraction is discussed. Next the *RAKE* algorithm is shown. The proposition of some modification of the original version of the *RAKE* method is the main purpose of the article. These changes should improve the quality of results obtained for text documents prepared in Polish. Also the exemplary application of the modified version of the algorithm is presented.

Keywords: keywords and keyphrases identification, keywords and keyphrases extraction, text mining, exploratory analysis of text documents.