Z badań nad systemem rafinacji sieciowej
Identyfikacja sentymentów

Włodzimierz Gogołek, Dariusz Jaruga

Rafinacja to proces oczyszczania i uszlachetniania substancji naturalnych lub produktów przemysłowych w celu nadania im odpowiedniej czystości, barwy, zapachu. Przytoczona definicja procesu rafinacji została zaczerpnięta ze *Słownika języka polskiego*[1]. Odzwierciedla ona również sposób, w jaki jest przeprowadzana rafinacja substancji, którą stanowią duże zasoby informacyjne – Big Data. Oczekiwanym efektem tego procesu są nowe informacje ukryte we wspomnianych zasobach.

W systemie rafinacji informacji (RI) substancją podlegającą obróbce są materiały źródłowe (materiały) w formie tekstowej lub audio pozyskiwane z sieci lub z dużych zbiorów informacji dostępnych offline – Big Data[2]. Finalnym efektem zastosowania RI jest wynik statystycznej analizy wyrażeń kluczowych i znajdujących się w ich okolicy sentymentów, czyli wyrażeń, które oddają emocje, a są zapisane w cyfrowym zasobie informacyjnym. Dzięki RI można wyłuskać informacje nowe, wartościowe, ukryte w treściach, na przykład oceny zjawiska społecznego (poparcie, zadowolenie, negacja)[3].

Przystępując do badania określonego zjawiska społecznego (które może być związane np. z biznesem, polityką, medycyną i innymi branżami), pierwszym krokiem jest ustalenie słów lub wyrażeń, które są związane z określeniem/nazwą badanego zjawiska. Takie słowo klucz, lub wyrażenie kluczowe z nim związane, jest w tej metodzie określone terminem „słup" (np. nazwa partii, firmy, nazwisko). Drugim krokiem będzie wyróżnienie sentymentów (swego rodzaju ocen), trzecim – obliczenie frekwencji obecności sentymentów wokół słupów. Ostatnim – pominiętym w artykule – interpretacja wyników (np. ocena popularności marki, jej ocena).

Zasygnalizowane etapy rafinacji stanowią przedmiot dalszej części wywodu. Dokumentują ważne ogniwo procesu rafinacji informacji sieciowych, jakim jest identyfikacja statystycznie istotnych sentymentów[4].

---

[1] M. Szymczak, *Słownik języka polskiego*, Warszawa, 1978.

[2] V. Marx, *The big challenges of Big Data* [w:] "Nature" 2013, vol. 498; W. Gogołek, P. Kuczma, *Rafinacja informacji sieciowych na przykładzie wyborów parlamentarnych,* „Studia Medioznawcze" 2013, nr 2 (53), s. 89.

[3] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *From Data mining to knowledge discovery in Database*, www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf [dostęp: 11.11.2016].

[4] Ch. Curme et al., *Quantifying the semantics of search behavior before stock market moves*, „PNAS" 2014, nr 32; J. Smailovič, *Predictive sentiment analysis of Tweets: A stock market application* [w:] *Human-computer interaction and knowledge discovery in complex, unstructured, Big Data* 2013, s. 77–88.

Nazwa zjawiska „słup"

Jak już wspomniano, w zależności od przedmiotu badań słupem może być nazwa marki, produktu, partii politycznej, organizacji, miasto, nazwisko osoby, np. polityka itp. Słup można traktować szerzej i nie musi być ograniczony tylko i wyłącznie do jednego słowa lub wyrażenia, lecz może stanowić cały zestaw słów i wyrażeń będących synonimami, antagonizmami lub zbiorem słów i wyrażeń dotyczących danego tematu (np. budżet, wynik ekonomiczny, raport, audyt etc.). Słup może obejmować wszelkie możliwe odmiany słowa lub wyrażenia przez osoby, czasy (przyszły, przeszły, teraźniejszy), tryby (przypuszczający, rozkazujący), imiesłowy, neologizmy, włącznie z uwzględnieniem słów zawierających błędy ortograficzne, literówki i coraz powszechniej stosowane hasztagi[5].

Przykładowo dla słowa „leczenie" słup może obejmować zestaw 38 wyrazów: leczą, lecząc, leczący, leczcie, leczę, leczmy, leczony, leczy, leczycie, leczyć, leczyli, leczyliby, leczylibyście, leczylibyśmy, leczyliście, leczyliśmy, leczył, leczyła, leczyłaby, leczyłabym, leczyłabyś, leczyłam, leczyłaś, leczyłby, leczyłbym, leczyłbyś, leczyłem, leczyłeś, leczyło, leczyłoby, leczyły, leczyłyby, leczyłybyście, leczyłybyśmy, leczyłyście, leczyłyśmy, leczymy, leczysz. Zbiór obejmuje odmianę czasownika przez osoby dla czasu teraźniejszego, przeszłego i przyszłego złożonego, odmianę wynikającą z trybu przypuszczającego, rozkazującego oraz imiesłowy.

Dobór słupa jest pierwszym krokiem, który należy wykonać przed przystąpieniem do wyznaczenia słów lub wyrażeń będących sentymentami.

Metodyka RI przewiduje, że słupy mogą być określone za pomocą trzech procedur: (1) dzięki intuicji badacza na podstawie przeglądu losowej próby tekstów ze zbioru źródłowego, na którym zostanie przeprowadzone badanie; (2) dostępnych słowników wyrazów (słowników)[6], które można uznać za sentyment (co dotychczas sprawdzano doświadczalnie); (3) na podstawie analizy częstotliwościowej wyrazów ze wskazanego zbioru źródłowego (AC). Opracowane narzędzie do liczenia częstotliwości pozwala również wyróżnić nietypowe wyrazy lub neologizmy, które funkcjonują w badanym zbiorze tekstów, a których częstotliwość występowania jest duża. W ten sposób określone słowa i wyrażenia są poddawane odmianie i docelowo stają się słupem. Zaletą narzędzia jest możliwość wykonania analizy na całym materiale źródłowym, a nie tylko na jego próbie losowej.

Na Wydziale Dziennikarstwa, Informacji i Bibliologii Uniwersytetu Warszawskiego zostało opracowane autorskie narzędzie, które oblicza częstotliwości wszystkich wyrazów znajdujących się w badanym materiale. Otrzymane w ten sposób wyniki na dalszym etapie – w celu identyfikacji sentymentów – są analizowane zgodnie z dwiema procedurami. Badacz wybiera odpowiednie słowa lub wyrażenia (sentymenty) spośród wykazanych słów (na podstawie AC lub/i słowników) do dalszego etapu (ACB) lub identyfikacja następuje automatycznie (ACA) na podstawie dalej opisanej procedury weryfikacji statystycznej istotności wyrazów uznawanych jako sentyment.

Sentymenty

Kolejnym krokiem jest znalezienie wyrażeń – sentymentów, które niosą określony ładunek emocjonalny (np. pozytywny, negatywny albo neutralny) i przez analizę frekwencji poszcze-

[5] Hasztag to pojedyncze słowo lub wyrażenie pisane bez spacji, poprzedzone symbolem #, np. #dziejesię. Hasztagi są stosowane w mikroblogach, serwisach społecznościowych, na stronach internetowych itp. Hasztag umożliwia grupowanie wiadomości. W chwili pisania niniejszej pracy hasło „hashtag" nie było dostępne w m.in. w *Encyklopedii Britanica* czy w *Encyklopedii PWN*. Dlatego na potrzeby niniejszego opracowania przyjęto definicję podaną w Wikipedii pod adresem https://pl.wikipedia.org/wiki/Hasztag [dostęp: 05.10.2016].

[6] W. Gogołek, P. Kuczma, *Rafinacja informacji sieciowych…,* dz. cyt.

Rys. 1. Łańcuch rafinacji zasobów sieciowych

Źródło: opracowanie własne

gólnych sentymentów, będących w sąsiedztwie słupa w zadanym przedziale czasu, pozwalają obliczyć np. stosunek społeczeństwa do badanego zagadnienia.

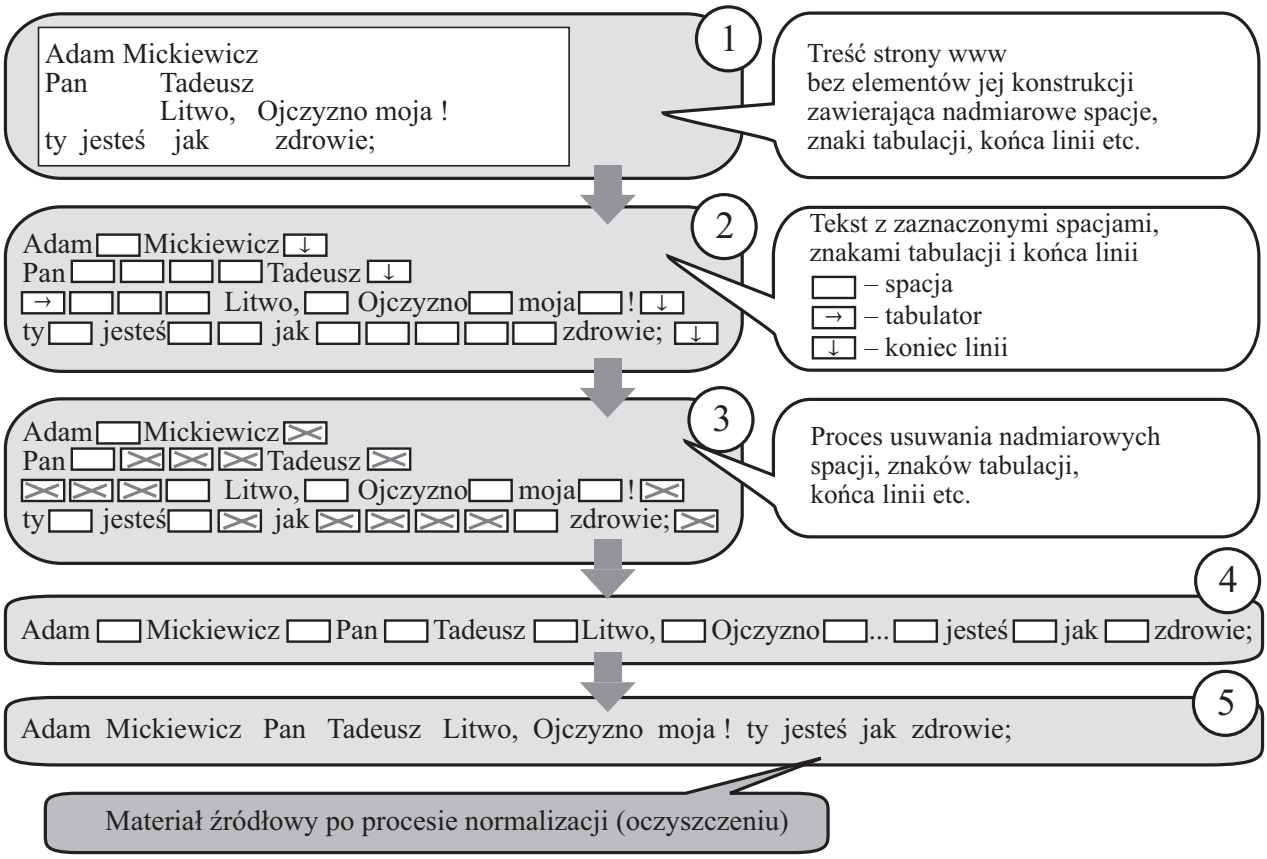
Sentymenty, podobnie jak słupy, mogą zostać wyznaczone na kilka sposobów, m.in. przez ręczną analizę losowo wybranych materiałów źródłowych, przez opracowane wcześniej słowniki słów i wyrażeń lub wspomniane autorskie narzędzie ACA opracowane na WDIB UW. Pozwala ono na przeprowadzenie analizy na całym materiale źródłowym, które będzie wykorzystane do dalszego badania – identyfikacji sentymentów.

Materiał źródłowy

Zanim zostanie uruchomiona procedura identyfikacji sentymentów, materiał źródłowy zebrany przez autorskiego robota BigData jest poddawany obróbce przygotowującej. Można ją podzielić na trzy etapy: odfiltrowanie z bazy danych robota treści podlegających badaniu; oczyszczenie danych i przekształcenie ich do postaci znormalizowanej wymaganej przez program; wykonanie właściwych obliczeń. Materiały dostępne w sieci ze względu na sposób prezentacji można podzielić na cztery podstawowe grupy: materiały tekstowe, obrazy, dźwięk i wideo.

Na obecnym etapie badań metoda wyznaczania sentymentów w okolicy słupów sprowadza się *de facto* do analizy materiałów czysto tekstowych. Nie wyklucza to wykorzystywania materiałów audio poddanych dostępnym narzędziom służącym do analizy dźwięków mowy ludzkiej. Dla materiałów graficznych zawierających teksty stosuje się standardy OCR (Optical Character Recognition).

Zgodnie z zasygnalizowanymi dalej etapami działania programu do obliczenia wyrazów-sentymentów pierwszym z nich jest odfiltrowanie rekordów przeznaczonych do badania z bazy danych robota BigData. To filtrowanie w głównej mierze sprowadza się do wyboru pewnego podzbioru treści zebranych przez BigData. Rekordy przeznaczone do badania mogą być wybierane na podstawie kilku kryteriów, takich jak przedziały czasu (data/godzina od–do), źródeł informacji, słów lub wyrażeń występujących w treści albo



Rys. 2. Proces normalizacji tekstu źródłowego

Źródło: opracowanie własne

w tytule. Wybrane w ten sposób dane są przekazywane do kolejnego etapu, w którym zostaną oczyszczone i przekształcone do postaci znormalizowanej wymaganej przez program. Czyszczenie strony internetowej oznacza w praktyce wykasowanie z jej źródła wszystkich znaczników – w wyniku tej operacji z materiału źródłowego pozostaje jedynie użyteczna treść dokumentu.

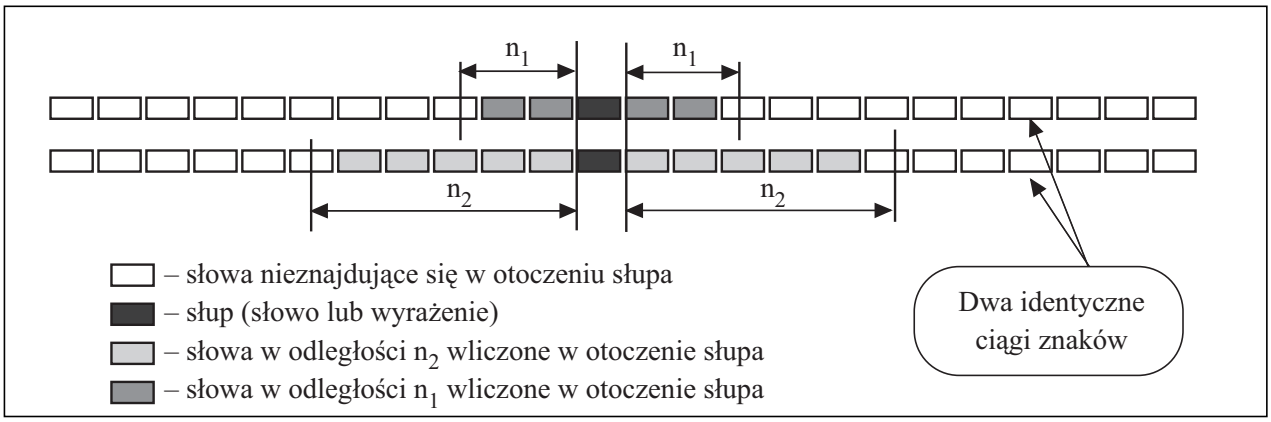
Pozyskany materiał tekstowy z dokumentu źródłowego, zanim zostanie poddany dalszej analizie, musi być doprowadzony do postaci znormalizowanej. Normalizacja polega na tym, że z tekstu są usuwane nadmiarowe znaki spacji, tabulacji i końca linii. W wyniku przeprowadzonej normalizacji tekstu (rys. 2) otrzymujemy jedną linię tekstu, w którym znajduje się cała treść dokumentu, a poszczególne wyrazy są od siebie oddalone nie więcej niż o jedną spację. W znormalizowanym tekście pozostawione są również znaki interpunkcyjne, nawiasy, itp.

Znormalizowana postać tekstu jest wymagana przez program do obliczania częstotliwości słów znajdujących się w okolicy – wcześniej opisanego słupa.

Frekwencje

Kolejnym krokiem po procesie normalizacji tekstu jest wykonanie procedur polegających na obliczeniu częstotliwości słów występujących w okolicy słupa, tj. oddalonych od niego o zadaną liczbę znaków „n”[7], nie większą niż określona wartość zdefiniowana jako parametr wejściowy.

[7] Na potrzeby niniejszego opracowania otoczenie słupa jest liczone w znakach i zostało oznaczone literą „n” lub w razie potrzeby literą „n” z indeksem dolnym, np. n_1, n_2 itp.



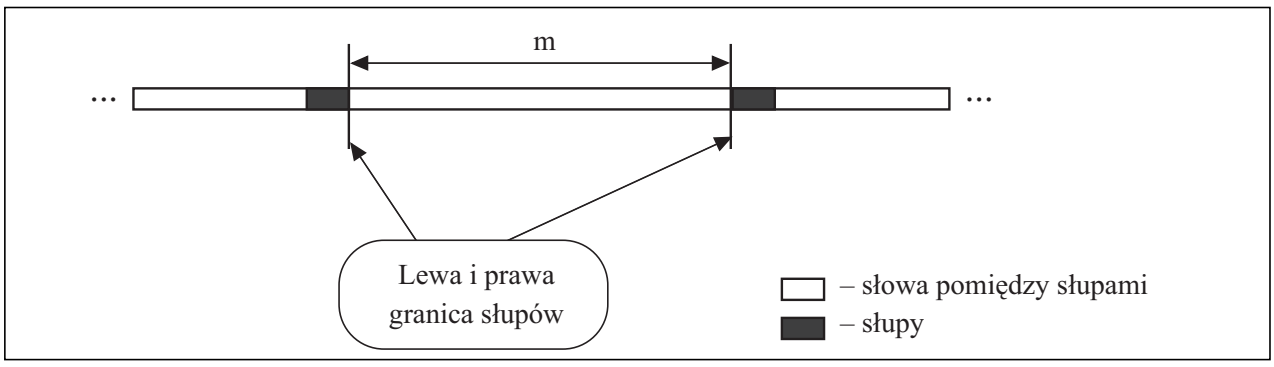
Rys. 3. Liczba słów zakwalifikowanych do otoczenia słupa w zależności od wartości parametru „n"

Źródło: opracowanie własne

Im większa wartość tej odległości, tym więcej słów zostanie zakwalifikowanych do grupy słów znajdujących się w otoczeniu słupa. Schematycznie przedstawiono to na rysunku 3.

W pierwszym przypadku dla „n_1" w otoczeniu słupa znalazły się cztery słowa, w drugim dla „n_2" w otoczeniu słupa jest 10 słów. W sytuacji, gdy granica „n" znaków wypadnie w środku słowa, a nie na jego granicy, takie słowo nie jest zaliczane do otoczenia słupa, nie ma tu znaczenia pomiędzy jakimi literami wewnątrz danego słowa przechodzi koniec wyznaczonego obszaru przez parametr „n". Jako granicę słowa/wyrazu należy rozumieć punkt styku spacji i pierwszej litery słowa lub ostatnią literę i następującą po niej spację.

Do otoczenia słupa są wliczane te słowa, które mieszczą się w całości w przedziale „n" znaków lub znajdują się na granicy tego obszaru – granicy słowa. Im wartość parametru „n" jest większa, tym więcej słów (będących podstawą identyfikacji sentymentów) zostanie zakwalifikowanych do otoczenia słupa. W praktyce parametr „n" jest ustawiany na wartość z przedziału od 10 do 60 znaków. Zatem parametr „n" w istotny sposób wpływa na zbiór słów znajdujących się w otoczeniu słupa, ale nie jest jedynym parametrem, który decyduje o tym, jakie słowa zostaną wliczone do otoczenia słupa. Na słowa należące do otoczenia słupa ma również wpływ wzajemne położenie dwóch słupów w badanym tekście względem siebie. Najbardziej istotną rzeczą jest tutaj wzajemna odległość pomiędzy dwoma słupami, która dla potrzeb niniejszego opracowania została określona parametrem „m". Innymi słowy mówiąc, „m" jest odległością w znakach pomiędzy jed-



Rys. 4. Ilustracja lewej i prawej granicy dwóch słupów i odległości pomiędzy nimi określonej parametrem „m"

Źródło: opracowanie własne

nym słupem i drugim, a dokładniej – pomiędzy lewostronną i prawostronną granicą słupa, która jest określana analogicznie jak dla słowa, co przedstawiono na rysunku 4.

W badanym znormalizowanym już tekście źródłowym liczba słupów może zawierać się od zera do pewnej skończonej liczby naturalnej. Nawet jeśli słup występuje w badanym tekście tylko raz, należy uwzględnić jego położenie względem początku i końca tekstu jako całości. Dla przypomnienia – znormalizowany tekst jest pojedynczą linią, gdzie w sposób jednoznacznie określony występuje jeden początek i koniec. Dla dwóch lub większej liczby słupów dodatkowo trzeba przeanalizować wzajemne położenie słupów względem siebie, ponieważ od tego zależy sposób zliczania słów znajdujących się w ich otoczeniu. W badaniach wyróżniono sześć kombinacji wzajemnego położenia różnych słupów względem siebie. Wyczerpują one możliwe sytuacje i tym samym dają podstawę, by w pełni opracować algorytm procedury zliczania częstotliwości słów w otoczeniu słupa.

Jeden z tych przypadków dotyczy sytuacji, w której dwa badane obszary otoczenia słupów zachodzą na siebie. Wówczas słowa znajdujące się zarówno w części wspólnej, jak i na jej dwóch granicach, muszą zostać policzone jednokrotnie. Tym samym, wszystkie słowa znajdujące się pomiędzy jednym a drugim słupem są wliczane do otoczenia.

Położenie dwóch lub większej liczby słupów względem siebie zawsze można rozbić na jeden z sześciu przypadków położenia dwóch słupów względem siebie. Upraszcza to również algorytm programu komputerowego do obliczania częstotliwości słów w otoczeniu słupa. Algorytm zliczania częstotliwości słów w otoczeniu słupa składa się z kilku kroków i zakłada, że przekazany do badania materiał źródłowy został już oczyszczony i znormalizowany.

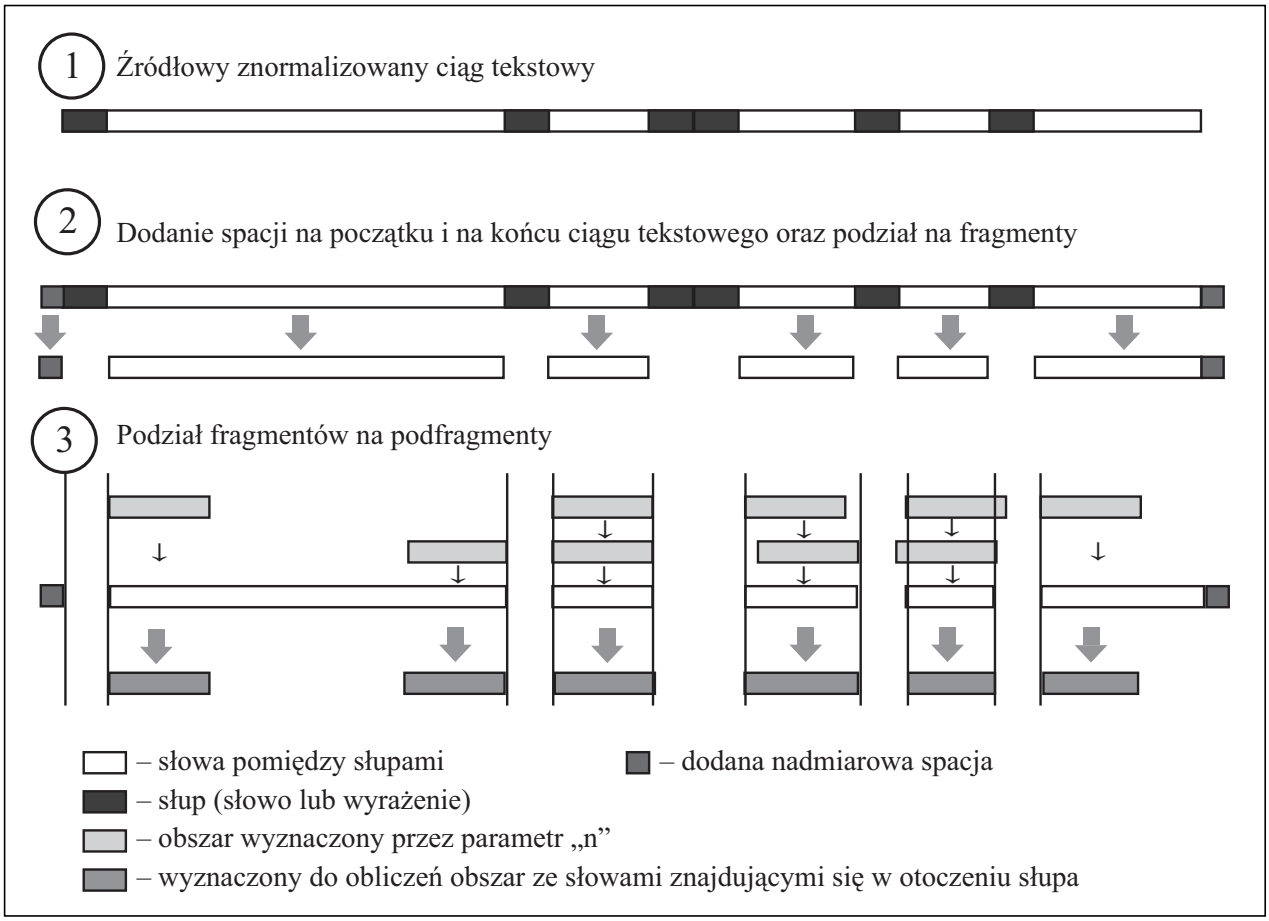
W pierwszym kroku do znormalizowanego ciągu tekstowego na jego początku i końcu jest dodawana pojedyncza spacja. Dzięki temu mamy pewność, że wszystkie słupy znajdą się w środku badanego ciągu tekstowego i eliminujemy dwa przypadki szczególne, gdy słup znajduje się dokładnie na początku lub na końcu badanego ciągu tekstowego, co sprowadzało by się do skomplikowania obliczania częstotliwości dla pierwszego i ostatniego fragmentu tekstu.

W drugim kroku znormalizowany ciąg tekstowy z dodanymi na początku i końcu spacjami należy podzielić na fragmenty, w których linią podziału tego tekstu jest słup. W ten sposób badany ciąg zostanie podzielony na części. Jeśli liczba słupów w tekście wynosi zero, algorytm na tym etapie kończy swoje działanie.

W trzecim kroku każdy fragment jest badany pod względem jego długości, czyli *de facto* pokazuje on, jaka jest rzeczywista odległość pomiędzy słupami w danym fragmencie tekstu. Następnie są obliczane częstotliwości wstępowania poszczególnych słów we fragmentach i następnie sumowane. Częstotliwości cząstkowe obliczone dla każdego z fragmentów – agregowane i sumowane. Po zsumowaniu informacji z każdego z fragmentów otrzymujemy pełną informację o częstotliwości wszystkich słów znajdujących się w otoczeniu słupa zdefiniowanym przez parametr „n". Schematycznie działanie algorytmu zostało przedstawione na rysunku 5.

W wyniku zastosowanej procedury zostaje utworzony zbiór frekwencji wszystkich wyrazów występujących w zadanym sąsiedztwie każdego słupa. Te dane stanowią podstawę wyróżnienia najczęściej występujących wyrazów, a wśród nich – tych, które są istotnymi (najczęściej występują) sentymentami[8].

[8] Y. Hongliang et al., *Identifying sentiment words using an optimization-based model without Seed Words,* https://www.google.pl/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0ahUKEwjEl5-W966XQAhUDjSwKHU_OAT4QFgguMAI&url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel7%2F7222492%2F7222827%2F07222836.pdf&usg=AFQjCNHT85uTAy_yYKjle7fngNTF67jctw [dostęp: 11.11.2016].



Rys. 5. Podział badanego tekstu na fragmenty

Źródło: opracowanie własne

Identyfikacja sentymentów

Jak już wspomniano, standardowo sentymenty są wyznaczane na podstawie badań subiektywnych ocen określonej grupy osób i dostępnych słowników. Ta metoda nie wykorzystuje w pełni potencjału, jaki drzemie w zasobach Big Data, tzn. korzystania z sentymentów identyfikowanych w czasie rzeczywistym. Służy temu ACA, czyli procedura samouczącej się identyfikacji sentymentów. Jako całkowicie innowacyjny produkt wymaga badań podstawowych (w przypadku tej procedury brak jakichkolwiek publikacji i informacji o wdrożeniach), zbierania doświadczeń (pracy na dużych zbiorach), które determinują trafność/istotność identyfikacji sentymentów.

W pierwszym kroku w procedurze ACA z bazy danych, w funkcji czasu, są wyróżniane zbiory najczęściej występujących wyrazów wokoło słupa (przedmiotem badań może być przykładowo firma, produkt, osoba itp.). Dla każdego z tych wyrazów jest liczona frekwencja (zmienna Wk) w interwałach czasowych (t_1, t_2…, t_n). Efektem będzie wyodrębnienie zbioru zmiennych Wk(t). Następnie w podobny sposób są pozyskiwane/obliczane oceny (np. wyniki badań rynkowych, CBOS, OBOP, praca eksperymentalnej wersji RI) dla wszystkich wyrazów (punkt odniesienia) sentymentów Sk(t) (k – nazwa, kierunek +/– oraz t – czas) w tych samych interwałach czasowych. Oceny (Sk) stanowią drugą zmienną. Wartość statystycznej istotności związku Sk(t) ze wszystkimi zmiennymi zbioru Wk(t) wskazuje zasadność wyboru poszukiwanych sentymentów. Wynikiem ACA są najistotniejsze statystycznie sentymenty dla wskazanego słupa w próbie ograniczonej wielkością analizowanych zbiorów (w praktyce są to miliony wpisów). Innymi słowy, sentymentem może się okazać wyraz, który

potocznie nie jest uznawany za niosący sobą przekaz emocjonalny. Np. podczas wstępnych badań związanych z frekwencjami potencjalnych sentymentów wobec słup = UCHODŹCY, są: Polska, Niemcy, Europa, UE, Merkel i Ukraina; słup = WAKACJE, są: zdjęcia, zapomnienie, komputer i smartfon; ZDROWIE: dochód, dieta, jedzenie, palenie, stres i alkoholizm.

Wynik obliczeń statystycznej zależności pomiędzy sentymentami a zjawiskiem (np. ocena osoby/produktu) w interwałach czasowych (t_1, t_2…, t_n) stanowi o możliwości predykcji na podstawie sentymentów (predyktorów) zjawiska w czasie $t_n + 1$ (poprawy lub pogorszenia wartości notowań/oceny). Te obliczenia wykorzystują analizę regresji wielokrotnej wykorzystanej do budowy modelu, który będzie możliwie jak najlepiej dopasowany do danych empirycznych w czasie przed t_n i pozwalał oszacować stan zjawiska w czasie $t_n + 1$. Wynika to z faktu, że do danych otrzymanych lepiej pasuje model regresji niż przypadek.

Zakończenie

W artykule – z oczywistych powodów – nie opisano detali prezentowanej procedury, której efektem jest możliwość oceny przebiegu zjawisk w przeszłości, w czasie rzeczywistym i ich predykcja, głównie dzięki identyfikacji sentymentów i ich poprawnej syntezie[9].

Badaniom nad omawianym przedmiotem towarzyszy szereg pobocznych problemów badawczych, np. trudności z jednoznacznym wyróżnieniem określenia nazwy przedmiotu badań, m.in. w badaniach wyborczych powstał problem (rozwiązany) z różnorodnymi określeniami partii PIS i PO[10].

Fundamentalnym warunkiem powodzenia RI jest dysponowanie dostateczną ilością informacji, by przekroczyć próg nieufności do uzyskanych wyników pozwalających na uzyskanie wiarygodnych wyników, np. predykcję zjawiska. System musi „się uczyć" – osiągać zakładaną statystyczną istotność. Rozwiązaniem (poza wielkością zbioru źródłowego) jest intensyfikacja prac nad zastosowaniem funkcji regresji wielokrotnej i uwzględniania w predykcji więcej niż jednego parametru – frekwencji sentymentów. Chodzi tutaj o łączenie zjawisk/przedmiotów badań, które są badane niezależnie, lecz wskazują wzajemne zależności.

Część zjawisk/przedmiotów badań jest zbyt losowa i na podstawie dotychczas dostępnych narzędzi/metod nie poddaje się analizie statystycznej prowadzącej do predykcji. Rozwiązaniem problemu będzie zdefiniowanie obszarów, dla których RI może być narzędziem szeroko rozumianej diagnozy (np. identyfikacja zagrożeń).

---

[9] M. Huberty, *Awaiting the second revolution: From digital noise to value creation,* http://eds-1a-1ebscohost-1com-1ebsco.han.buw.uw.edu.pl/eds/detail/detail?vid=3&sid=16298e68-0990-4883-9fc20e0e51a4dab5%40sessionmgr4004&hid=4205&bdata=Jmxhbmc9cGwmc2l0ZT1lZHMtbGl2ZSZzY29wZT1zaXRl#db=edb&AN=10151-6526, [dostęp: 06.05.2015]; Y. Liu, *Big Data and predictive business analytics,* „Journal of Business Forecasting" 2015, http://eds-1a-1ebscohost-1com-1ebsco.han.buw.uw.edu.pl/eds/pdfviewer/pdfviewer?sid=16298e68-0990-4883-9fc2-0e0e51a4dab5%40sessionmgr4004&vid=6&hid=4205 [dostęp: 05.05.2015].

[10] W. Gogołek, P. Kuczma, *Rafinacja informacji sieciowych…,* dz. cyt.

Z badań nad systemem rafinacji sieciowej. Identyfikacja sentymentów

From research on the system of refining the Web. Identifying sentiment words

Włodzimierz Gogołek, Dariusz Jaruga

SŁOWA KLUCZOWE
informacja, internet, Big Data, kolekcjonowanie informacji, identyfikacja sentymentów, analiza sentymentów

STRESZCZENIE
Dostępny potencjał mocy obliczeniowych i pamięci komputerowych stworzył niedostępne wcześniej warunki do analizy dużych zasobów informacyjnych – Big Data. W procesie tej analizy można wykorzystywać procedury kolekcjonowania informacji i ich analizy do trafnej oceny – w kategoriach emocjonalnych (sentymentów – dobry, zły) badanych zjawisk w przeszłości, w czasie rzeczywistym, a także do predykcji. Artykuł jest prezentacją kluczowej części tej procedury – istoty automatyzacji procesu identyfikacji sentymentów.

KEY WORDS
information, internet, Big Data, collecting information, identifying sentiment words, sentiment analysis

ABSTRACT
Available potential of computing power and computer memory had created, previously unavailable conditions for the analysis of, large information resources – Big Data. In the process of this analysis can be used procedures for collecting information and analysis for the accurate assessment – in terms of emotional (sentiments – good, bad) of studied phenomena – in the past, in real time, as well as to the prediction. The article is a presentation of the key parts of this procedure – being automate the process of identifying sentiment words.

studia medioznawcze
NUMER 4 (67) 2016
ISSN 1641-0920
ISSN 2451-1617