Fides, Ratio et Patria. Studia Toruńskie 3/2015
ISSN 2392-1463

Tomasz Kowalski[*], **Mateusz Molek**[**]
Wyższa Szkoła Kultury Społecznej i Medialnej w Toruniu

Analiza doniesień prasowych z wykorzystaniem technik przetwarzania języka naturalnego

1. Użytkowanie portali informacyjnych

W polskojęzycznym Internecie nietrudno jest znaleźć strony internetowe mające charakter informacyjny – ogólny lub dziedzinowy, a wśród najpopularniejszych polskojezycznych stron internetowych[1] można odnaleźć zarówno takie, które publikują informacje pozyskane z własnych źródeł[2], jak i takie, które wprost odsyłają do treści pochodzących z innych źródeł[3].

W przypadku tych pierwszych można mieć nadzieję, że konkurują między sobą pod względem tego, kto pierwszy podejmie daną sprawę i jaki poziom szczegółowości jej przedstawienia zaprezentuje. Taka unikatowa informacja zdaje się, z czytelniczego punktu widzenia, najbardziej wartościowa. Oferowane są również informacje „z drugiej ręki", mogące przyjąć postać przedruków, parafraz czy cytowań. Taka zduplikowana, a więc dostępna na wielu portalach, informacja (np. rys. 1) może być – z perspektywy odbiorcy – postrzegana jako wartościowa (gdyż nie trzeba jej szukać w innych miejscach), ale może też przynosić szkodę.

[*] **Tomasz Kowalski** – doktor nauk matematycznych w zakresie informatyki, wykładowca WSKSiM; naukowo zajmuje się zagadnieniami analizy tekstu, programowania rozproszonego, sieci komputerowych i bezpieczeństwa.

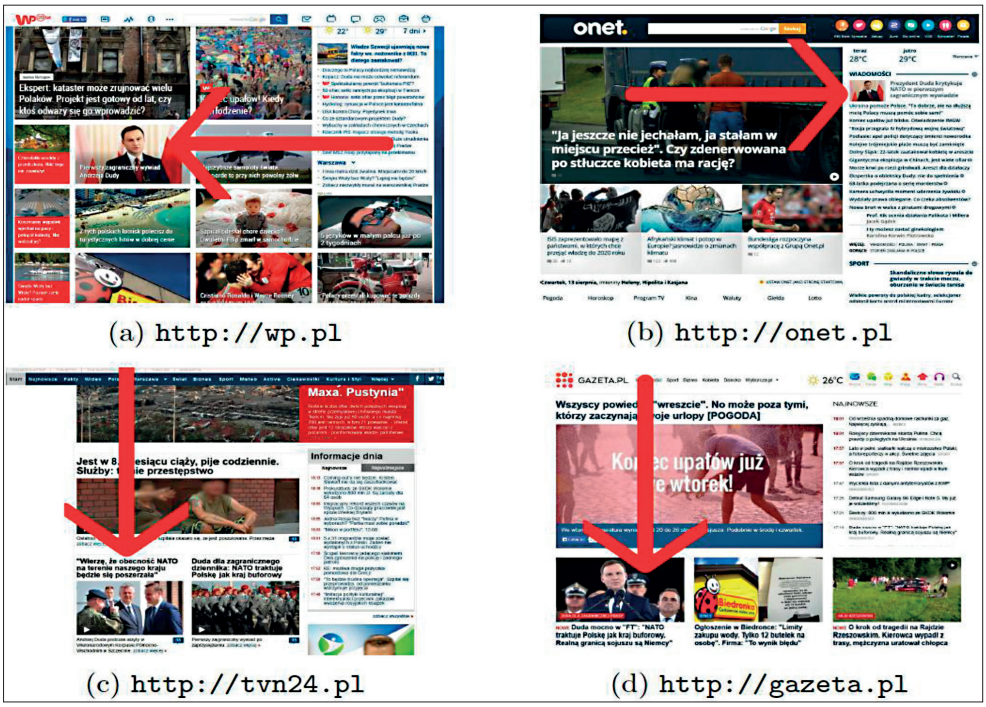
[**] **Mateusz Molek** – inżynier (informatyka – techniki multimedialne), absolwent WSKSiM; kontynuuje studia drugiego stopnia na UMK (informatyka – programowanie).

[1] Istnieje wiele publicznie dostępnych rankingów popularności stron internetowych (Alexa: http://www.alexa.com/topsites, SimilarWeb: http://www.similarweb.com/global, itp.). Dokładna pozycja strony w różnych rankingach może się różnić w zależności od użytej metodologii. Najpowszechniejszymi czynnikami decydującymi o pozycji w rankingu są liczba wyświetleń i liczba użytkowników.

[2] Prasowe, radiowe lub telewizyjne agencje informacyjne publikują treści również na swoich stronach internetowych.

[3] Przykładem popularnej polskojęzycznej witryny agregującej treści pochodzące z innych portali może być http://o2.pl. Na stronie głównej prezentowane są wpisy m.in. ze stron takich, jak: sfora.pl, pudelek.pl, money.pl, biztok.pl czy wp.pl.

Lektura czegoś, co poznało się już wcześniej, szczególnie w kontekście bieżących doniesień medialnych, nie przynosi korzyści i bez wątpienia może zostać potraktowana jako tzw. narzut użytkowy. Polega on w tym przypadku na obciążeniu użytkownika (czytelnika) koniecznością kognitywnego wysiłku (rozpoznanie aktywnych elementów strony internetowej wśród bogatej oprawy graficznej, czytanie, ruch myszką lub kliknięcie w celu wyświetlenia streszczenia lub otwarcia pełnego artykułu itp.) tylko po to, aby po krótkiej chwili mógł sobie uświadomić, że nic nie zyskał (przypomniawszy sobie, że czytał już wcześniej to samo lub coś bardzo podobnego). Redukowanie wszelkiego rodzaju tzw. narzutów użytkowych powinno być priorytetem przy projektowaniu interakcji[4].



Rys. 1. Przykład tej samej informacji dostępnej na wielu portalach. Strzałkami zaznaczono „zwiastuny" artykułów z 13 sierpnia br., dotyczących wywiadu prezydenta Andrzeja Dudy dla „Financial Times". Artykuły nieznacznie różnią się tytułami i treścią (źródło: własne)

Znaczny narzut użytkowy (mierzony poświęconym czasem i wysiłkiem intelektualnym) generowany przez zduplikowane informacje jest stanem faktycznym w Internecie, gdzie publikowanie „kopii" wiąże się z niemal zaniedbywalnym

[4] A. Cooper, R. Reimann, D. Cronin, *About Face 3: The Essentials of Interaction Design 3rd Edition*, Indianapolis 2007.

(w porównaniu z kosztem druku lub emisji radiowej czy telewizyjnej) kosztem[5]. Co więcej, tam gdzie przychód np. z reklam prezentowanych większej liczbie użytkowników przewyższy koszt infrastruktury do publikowania i jej obsługi, może być rynkowo zasadne powielanie każdej możliwej informacji[6].

Czytelnik, który bezwzględnie zaufa redakcji i autorom jednego portalu (w zakresie selekcji, szczegółowości i sposobu przedstawienia treści), zapewne uzna jego walory użytkowe i informacyjne za wystarczające. Każdy inny, z konieczności, musi się stać użytkownikiem wielu portali i (w typowej sytuacji) ponieść koszt (czas i wysiłek) związany z natłokiem treści w postaci zduplikowanych informacji.

2. Systematyczne śledzenie zmian treści na wielu portalach

Strony internetowe, których właściciele czerpią zyski z reklam na nich prezentowanych, są optymalizowane na maksymalizację czasu, jaki użytkownik na nich spędza, i liczbę wyświetleń. Niestety, nie idzie to w parze ze zwiększeniem czytelności i dostępności[7], i jeśli nie korzysta się z tzw. kanałów informacyjnych, zmniejszone walory użytkowe czynią śledzenie zmian na wielu portalach (a szczególnie systematyczne monitorowanie) wyjątkowo uciążliwym.

Współczesne strony, wśród różnych form prezentacji zawartych na nich treści (np. dla urządzeń o różnych rozmiarach ekranu lub poszczególnych użytkowników), oferują dostęp w postaci internetowych kanałów informacyjnych (ang. *web feed*). Techniczne taki kanał oznacza plik o ustalonym formacie, który jest systematycznie aktualizowany tak, aby jego zawartość odpowiadała nowym lub zmieniającym się treściom na stronie internetowej.

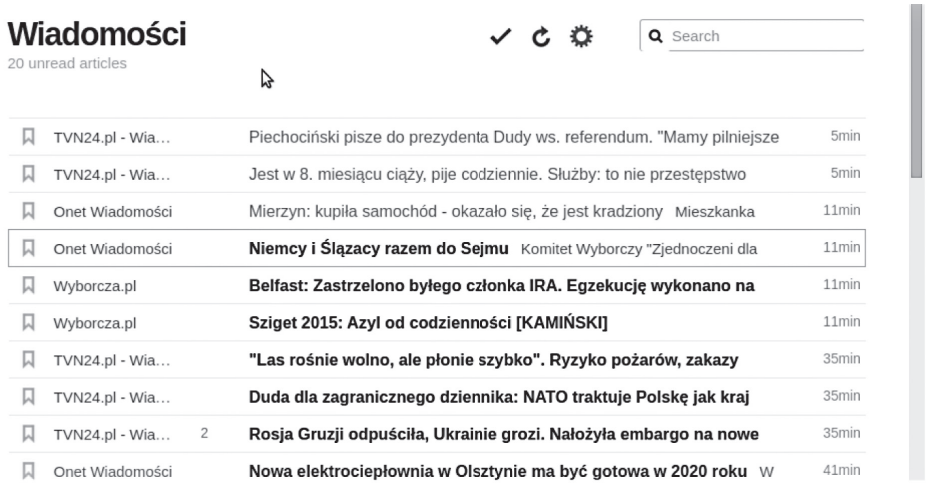
Czytelnik, pragnąc śledzić nowości na wielu portalach, może skorzystać z tzw. czytnika kanałów informacyjnych i wskazać odpowiednie pliki na poszczególnych stronach. Czytnik (np. rys. 2) jest aplikacją, która agreguje treści ze wszystkich źródeł i prezentuje je w kolejności chronologicznej, z zadanym poziomem szczegółowości (np. tylko tytuły, ze zdjęciem przewodnim, z pełnym tekstem).

---

[5] O koszcie publikowania w Internecie może świadczyć to, że jest wielu usługodawców, którzy bez opłat udostępniają swoją infrastrukturę (zaplecze sieciowe, serwerowe i programowe) każdemu chętnemu do publikowania treści. Przykładem jest tu każdy serwis społecznościowy lub serwis blogowy.

[6] Powszechność duplikowania informacji w Internecie jest szczególnie widoczna, gdy powielenie treści nie wiąże się z opłatami licencyjnymi lub autorskimi.

[7] Użytkownik odwiedza portal w poszukiwaniu pewnych treści i choć, dla własnej korzyści, można mu to utrudnić, to nie można go zupełnie pozbawić możliwości osiągnięcia tego celu. Porównując walory użytkowe wielu portali, można zauważyć, że do treści na nich dostępnych dużo łatwiej jest dotrzeć za pomocą urządzeń mobilnych. Zmniejszona przestrzeń ekranowa wymusza rezygnację z części oprawy graficznej, wprowadza ułatwienia w nawigacji i zmniejszenia liczby reklam.



Rys. 2. Przykładowy widok wpisów w czytniku kanałów informacyjnych (https://feedly.com). Najpowszechniejszy sposób prezentacji treści pochodzących z różnych źródeł polega na ich uporządkowaniu w kolejności publikacji – od najnowszych u góry. W każdym wierszu widoczne jest źródło artykułu, jego tytuł oraz fragment treści lub opisu. Tutaj trzy najnowsze wpisy zostały oznaczone jako przeczytane; nie pojawią się przy kolejnym wyświetleniu tego widoku, gdyż użytkownik zapoznał się z ich treścią lub na podstawie tytułu bądź opisu postanowił ich nie otwierać. Czytnik może przekierować użytkownika bezpośrednio na tę stronę portalu, na której znajduje się wskazany artykuł (źródło: własne)

Niektóre czytniki kanałów informacyjnych wskazują stopień popularności poszczególnych publikacji (na podstawie liczby użytkowników, którzy powiadomili lub udostępnili innym daną publikację), jednak nie oferują funkcjonalności umożliwiających wykrywanie i oznaczanie (potencjalnie) zduplikowanych treści.

Obecne czytniki kanałów nastawione są silnie na pełnienie funkcji osobistych organizatorów treści. Choć posiadają funkcje wspierające dla sieci społecznościowych (a więc udogodnienia dla udostępniania lub powiadamiania), często oferują również możliwość zapisywania (dla późniejszej lektury), dowolnego oznaczania pojedynczych wpisów (tagowania), a przede wszystkim wyszukiwania pełnotekstowego.

3. Określanie podobieństwa dokumentów tekstowych

Porównywanie tekstów w poszukiwaniu podobieństw i różnic jest jednym z zagadnień szeroko pojętego rozumienia i generowania języka naturalnego przez komputer[8]. Mówi się o nim często w kontekście tzw. wyszukiwania i ekstrakcji

[8] Przetwarzanie języka naturalnego (ang. *natural language processing*) to interdyscyplinarna dziedzina łącząca informatykę, sztuczną inteligencję oraz lingwistykę. Swoim zakresem

cech. Znanych jest wiele metryk umożliwiających numeryczne określenie tego, w jakim stopniu dwa dokumenty są podobne do siebie[9], jednak bez wątpienia najpopularniejszą z nich (gdyż szeroko stosowaną od wielu lat np. w modelowaniu preferencji[10], wykrywaniu trendów[11] czy rekomendowaniu[12]), jest tf-ifd (od ang. *term frequency–inverse document frequency*). Metoda ta polega, jak wskazuje nazwa, na określeniu częstości występowania tzw. termów w każdym dokumencie z osobna i w całej analizowanej kolekcji.

Ideowo porównywanie wzajemnego podobieństwa dokumentów (w pewnej ich kolekcji) z użyciem tf-ifd można opisać jako dwuetapowy proces, w którym najpierw dla każdego z dokumentów określa się, jak znaczące (w badanym dokumencie i w relacji do wszystkich dokumentów) jest każde ze słów, a następnie tak uzyskane charakterystyki każdego dokumentu zestawia się ze sobą parami. Słowo jest tym „ważniejsze" (charakterystyczne, specyficzne) dla danego dokumentu, im częściej w nim występuje, pod warunkiem że nie występuje równie powszechnie w innych dokumentach. Wysoki wynik mogą uzyskać w szczególności terminy techniczne, specyficzne określenia, nazwy przedmiotów, miejsc lub określenia ludzi.

Wywiad Andrzeja Dudy dla "Financial Times": NATO nie zauważyło przejścia Polski ze Wschodu na Zachód (a) `http://wp.pl`	Andrzej Duda dla "Financial Times": NATO traktuje Polskę jak kraj buforowy (b) `http://onet.pl`
Duda dla zagranicznego dziennika: NATO traktuje Polskę jak kraj buforowy (c) `http://tvn24.pl`	Duda mocno w "FT": "NATO traktuje Polskę jak kraj buforowy. Realną granicą sojuszu są Niemcy" (d) `http://gazeta.pl`

Rys. 3. Tytuły artykułów z rys. 1 posłużą zilustrowaniu obliczania podobieństwa z użyciem tzw. tf–idf (źródło: własne)

obejmuje między innymi rozumienie i generowanie tekstu oraz mowy, automatyczne tłumaczenia, generowanie streszczeń, a także wyszukiwanie i ekstrakcję cech.

9 Ch. Manning, P. Raghavan i H. Schütze, *An Introduction to Information Retrieval*, Cambridge 2009.

10 D. Billsus, M.J. Pazzani, *A Hybrid User Model for News Story Classification* w: *UM99 User Modeling. CISM International Center for Mechanical Sciences*, t. 407, ed. by J. Kay, Wien 1999, s. 99–108.

11 S. Phuvipadawat, T. Murata, *Breaking News Detection and Tracking in Twitter* w: *Web Intelligence and Intelligent Agent Technology*, t. 3, Toronto 2010, s. 120–123.

12 A. Elbadrawy, G. Karypis, *User-Specific Feature-Based Similarity Models for Top–n Recommendation of New Items* w: *ACM Trans.* Intell. Syst. Technol., t. 3 (6), New York 2015, s. 33:1–33:20.

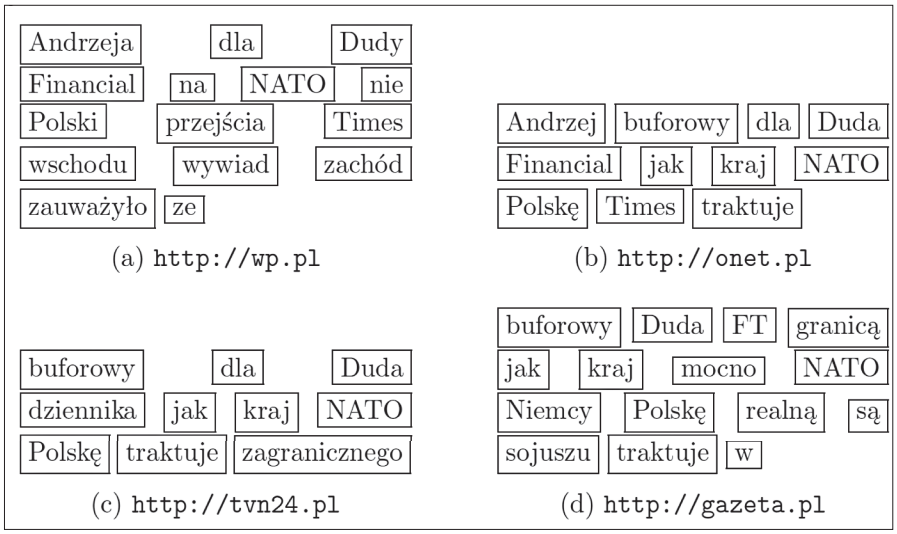
Tymczasem słowa powszechnie stosowane w danym języku, a więc te, które względnie często występują w każdym dokumencie, uznawane są za mniej istotne. Dokumenty traktujące o tych samych osobach, miejscach, przedmiotach itd. będą więc uznawane za podobne tym bardziej, im więcej będą miały tych samych znaczących słów.

Należy zauważyć, że przy zastosowaniu tej metody tekst potraktowany zostaje jedynie jako wielozbiór słów, co może wiązać się z utratą informacji zawartej np. w strukturze gramatycznej zdań i ich kolejności występowania w tekście. Przykład takiej reprezentacji dla zdań z rys. 3 przedstawiono na rys. 4.

Niech $\Omega = [t_1, t_2, ..., t_n]$ będzie wektorem wszystkich tzw. termów występujących w całej badanej kolekcji dokumentów. Wtedy dokument *d* możemy sformalizować w postaci wektora $[w_{t_1}, w_{t_2}, ..., w_{t_n}]$, gdzie w_{t_1} to waga powiązana z *i*–tym termem *t*. Przykład takiej reprezentacji dla zdań z rys. 4 przedstawiono na rys. 5.

Trudno ściśle określić, czym jest term. Intuicyjnie można traktować go jako słowo, a praktycznie często oznacza on tzw. token, a więc (w przybliżeniu) ciąg znaków oddzielony białymi znakami od innych tego typu ciągów. To, czy jako termy traktowane są znaki przestankowe, cyfry lub liczby, daty lub tzw. byty nazwane (ang. *named entities*), zależne jest od konkretnego zastosowania.

Waga termu jest nieujemną liczbą rzeczywistą równą iloczynowi $tf_{t,d} * idf_{t,D}$, gdzie $tf_{t,d}$ to tzw. częstości termu *t* w dokumencie *d*, a $idf_{t,D}$ to tzw. odwrócona częstość termu *t* w kolekcji $_D$.



Rys. 4. Przykład jednej z możliwych reprezentacji tytułów artykułów z rys. 3 w postaci „stokenizowanej". Pominięte zostały odstępy, cudzysłowy, kropki i dwukropki. Z wyjątkiem nazw własnych tokeny zapisano małymi literami. Przyjęto, że tokeny będą w kolejności alfabetycznej (źródło: własne)

Niech $f(t,d)$ oznacza stosunek liczby wystąpień terminu t w dokumencie d do liczby wszystkich termów w tym dokumencie. Wtedy częstość termu w dokumencie $tf_{t,d}$ określa się, w zależności od zastosowania, jako:

- bit informacji oznaczający fakty wystąpienia termu w dokumencie (a więc {0,1}),
- 1+ log(f(t,d)) lub 0, gdy term nie występuje w dokumencie,
- $K + (K - 1) = \frac{f(t,d)}{max f(t,d): t \epsilon D}$, gdzie K to stała normalizacyjna (najczęściej 0,5).

Ostatnia możliwość jest szczególnie przydatna dla uniknięcia faworyzowania dłuższych dokumentów.

Odwrotną częstość termu w kolekcji dokumentów $idf_{t,D}$ (gdzie D to zbiór wszystkich dokumentów w kolekcji) definiuje się jako logarytmicznie skalowany iloraz rozmiaru kolekcji do liczby dokumentów zawierający danych term, a więc $\log\left(1 + \frac{|D|}{|d \epsilon D, t \epsilon d|}\right)$.

	Andrzej	Andrzeja	buforowy	dla	Duda	Dudy	dziennika	Financial	FT	granicą	jak	kraj	mocno	na	NATO	nie	Niemcy	Polskę	Polski	przejścia	realną	są	sojuszu	Times	traktuje	w	wschodu	wywiad	zachód	zagranicznego	zauważyło	ze
(a)		*		*		*	*		*		*		*						*	*	*						*	*	*		*	*
(b)	*		*			*	*			*	*	*	*			*	*								*							
(c)			*				*			*		*	*	*		*	*								*					*		
(d)			*		*		*	*		*		*			*	*	*	*					*	*	*	*	*						

Rys. 5. Tytuły artykułów z rys. 4 zapisane w postaci 32-elementowych ciągów. Oznaczono te elementy ciągów, które wypełnione zostaną wagami tf–idf. Pozostałe elementy odpowiadają tokenom, które nie występują w danym zdaniu. Zostanie im przyporządkowana wartość 0 (źródło: własne)

Podobieństwo dwóch dokumentów określa się jako iloczyn skalarny ich wektorowych reprezentacji. Rysunek 6 przedstawia wektory zdań z rys. 5 wypełnione wagami, a rys. 7 – macierz podobieństwa badanych zdań.

4. Analiza podobieństwa doniesień prasowych

Dla zbadania możliwości zastosowania przedstawionych wcześniej metod do doniesień prasowych, na podstawie aktualnego rankingu Alexa[13], spośród

[13] Z aktualnym rankingiem Alexa można zapoznać się na stronie: http://www.alexa.com/topsites/countries/PL.

25 najpopularniejszych stron internetowych wybranych zostało sześć portali, które można uznać za ogólnoinformacyjne. Są to (w kolejności alfabetycznej): http://fakt.pl, http://gazeta.pl, http://interia.pl, http://onet.pl, http://tvn24.pl, http://wp.pl.

Niektóre wybrane do eksperymentu witryny oferują więcej niż jeden kanał informacyjny, np. zawierający tylko doniesienia z kraju, ze świata, gospodarcze, lokalne itd. W takiej sytuacji wybrany został jeden kanał, który można uznać za najbardziej ogólny.

$$[0., 0.317, 0., 0.317, 0., 0.250, 0.165, 0., 0.317, 0., 0.250, 0., 0.202, 0., 0., 0., 0., 0., 0.317, 0.317, 0., 0., 0., 0., 0., 0., 0.317, 0.317, 0., 0., 0., 0.317]$$

(a) `http://wp.pl`

$$[0.449, 0., 0.286, 0., 0., 0.354, 0.234, 0., 0., 0., 0.354, 0.286, 0.286, 0., 0., 0.286, 0.286, 0., 0., 0., 0., 0., 0., 0., 0., 0.286, 0., 0., 0., 0., 0., 0., 0., 0.,]$$

(b) `http://onet.pl`

$$[0., 0., 0.293, 0., 0., 0., 0.240, 0., 0., 0., 0., 0.293, 0.293, 0.460, 0., 0.293, 0.293, 0., 0., 0., 0., 0., 0., 0., 0.293, 0., 0., 0., 0., 0.460, 0., 0.,]$$

(c) `http://tvn24.pl`

$$[0., 0., 0.254, 0., 0.398, 0., 0.207, 0.398, 0., 0., 0., 0.254, 0., 0., 0., 0.254, 0.254, 0.398, 0., 0., 0., 0., 0.398, 0., 0.254, 0., 0., 0., 0., 0., 0., 0.,]$$

(d) `http://gazeta.pl`

Rys. 6. Wypełniona wagami tf–idf reprezentacja tytułów artykułów z rys. 5. Ponieważ w poszczególnych zdaniach tokeny nie powtarzają się, a więc częstości termów w danym zdaniu są takie same, wagi zależą w znacznej mierze od częstości termów w kolekcji (źródło: własne)

Do eksperymentu użyto następujących kanałów informacyjnych:

- http://www.fakt.pl/rss/rss.asp?k=18,
- http://serwisy.gazeta.pl/pub/rss/najnowsze_wyborcza.xml,
- http://kanaly.rss.interia.pl/fakty.xml,
- http://wiadomosci.onet.pl/11,kategoria.rss,
- http://www.tvn24.pl/najnowsze.xml,
- http://wiadomosci.wp.pl/rss.xml.

Syndykacja treści, a więc dostęp za pośrednictwem wymienionych kanałów, odbyła się w ciągu jednego tygodnia. Plik kanału informacyjnego pobierany był co godzinę, bez polegania na mechanizmie ETag[14].

	(a)	(b)	(c)	(d)
(a) Wywiad Andrzeja Dudy dla "Financial Times": NATO nie zauważyło przejścia Polski ze Wschodu na Zachód	1.	0.27409	0.09935	0.03441
(b) Andrzej Duda dla "Financial Times": NATO traktuje Polskę jak kraj buforowy	0.27409	1.	0.56168	0.41282
(c) Duda dla zagranicznego dziennika: NATO traktuje Polskę jak kraj buforowy	0.09935	0.56168	1.	0.42332
(d) Duda mocno w "FT": "NATO traktuje Polskę jak kraj buforowy. Realną granicą sojuszu są Niemcy"	0.03441	0.41282	0.42332	1.

Rys. 7. Wzajemne podobieństwo tytułów artykułów z rys. 6 obliczone na podstawie wag tf–idf

Kanał informacyjny może przyjąć postać RSS 0.9x[15], RSS 1.0[16], RSS 2.0[17], Atom 0.3 lub Atom 1.0[18]. Każdy z tych formatów jest w istocie plikiem XML[19], ale ze względu na różne struktury dokumentów i nazw elementów do dalszego przetwarzania wygodniej jest użyć biblioteki programistycznej, która ujednolici dostęp do kanałów, niż wprost parsera XML. Przykładem takiej biblioteki jest *feedparser*[20] dla *Python*[21].

[14] ETag (skrót od ang. *entity tag*) to jedna ze składowych protokołu HTTP, która wspiera walidację cache i umożliwia klientowi formułowanie warunkowych zapytań. W efekcie serwer może nie wysłać pełnej odpowiedzi, pod warunkiem że zawartość (np. pliku) się nie zmieniła.

[15] RSS 0.9x jest specyfikowany przez http://www.rssboard.org/rss-0-9-0 i http://www.rssboard.org/rss-0-9-1-netscape.

[16] RSS 1.0 jest specyfikowany przez http://web.resource.org/rss/1.0/.

[17] RSS 2.0 jest specyfikowany przez http://www.rssboard.org/rss-specification.

[18] Atom jest specyfikowany przez https://tools.ietf.org/html/rfc4287 i https://tools.ietf.org/html/rfc5023.

[19] XML to otwarty tekstowy format zapisu danych oparty na znacznikach. W założeniu ma być on czytelny zarówno dla człowieka, jak i dla maszyny. Jest standaryzowany przez World Wide Web Consortium w wersjach 1.0 (http://www.w3.org/TR/2008/REC-xml-20081126/) i 1.1 (http://www.w3.org/TR/2006/REC-xml11-20060816/).

[20] Feedparser jest otwartoźródłową biblioteką oferującą warstwę abstrakcji, zapewniającą unifikację dat, liberalną obsługę kodowania znaków i częściową kompatybilność kanałów RSS i Atom. Jest dostępna na https://pypi.python.org/pypi/feedparser.

[21] Python (https://www.python.org/) to otwartoźródłowy, interpretowalny, zorientowany obiektowo (oraz częściowo funkcyjny), rozszerzalny, niezależny od platformy język programowania.

Plik kanału informacyjnego, w uproszczeniu, składa się z arbitralnej (zmiennej w zależności od dostawcy kanału) liczby wpisów zawierających (w założeniu):

- tytuł artykułu, którego dotyczy dany wpis,
- opis będący streszczeniem, fragmentem lub całym artykułem; długość fragmentu jest arbitralna,
- datę publikacji lub aktualizacji artykułu,
- adres URL, zwykle prowadzący wprost do tej konkretnej strony na portalu, która zawiera artykuł.

W przypadku wymienionych kanałów liczba wpisów wynosiła od 10 do 50. Należy jednak zauważyć, że plik kanału ma zawsze tę samą długość (biorąc pod uwagę liczbę wpisów). Oznacza to, że aktualizowany jest w ten sposób, iż w przypadku pojawienia się nowszego wpisu pewien wpis (być może najstarszy) jest usuwany. W konsekwencji poleganie na mechanizmach wspierających obsługę aktualizowanych treści (np. wspomniany wcześniej ETag w HTTP) nie zabezpiecza przed zaobserwowaniem tych samych wpisów podczas kolejnych dostępów do pliku na serwerze.

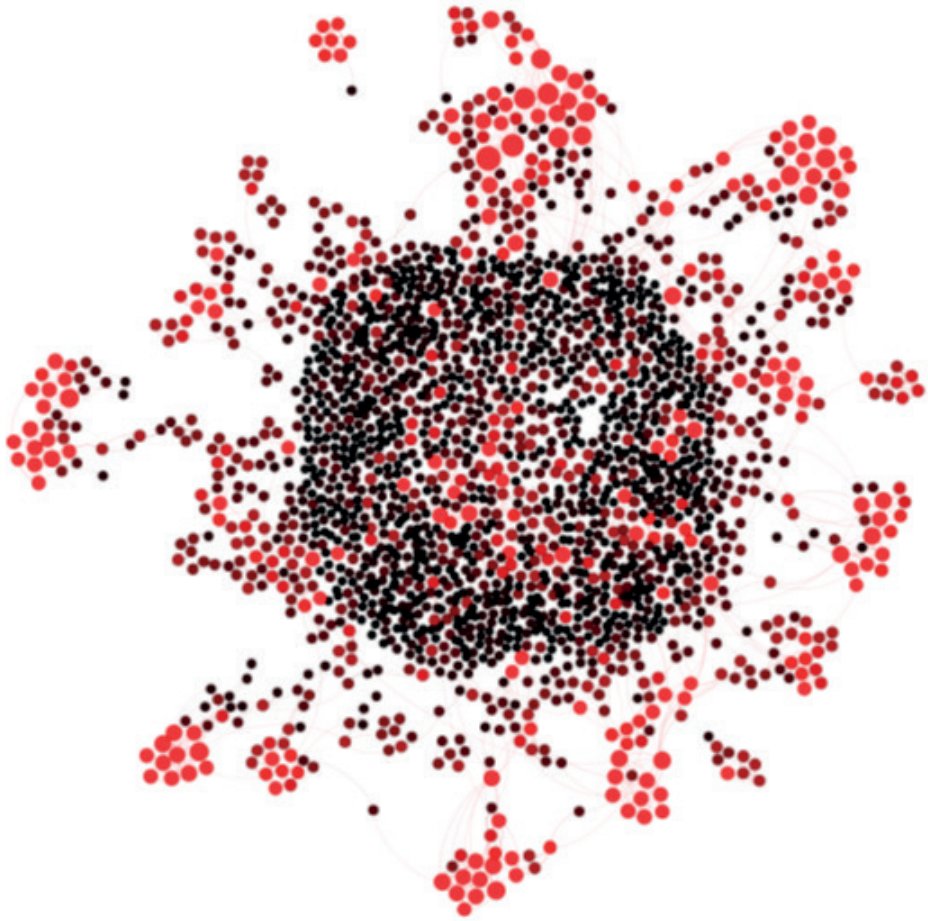
Dla dalszego przetwarzania konieczne było wyodrębnienie tytułu i tekstu. Powtórzenia wpisów, wynikające z możliwości aktualizacji jedynie części wpisów w kanale informacyjnym, wyeliminowano na podstawie adresów URL, do których każdy z wpisów odsyła.

Mając na uwadze, że tytuły artykułów są pojedynczymi zdaniami, a ich opisy dostępne w kanale są zwykle bardzo krótkimi fragmentami artykułów, postanowiono wszystkie wyrazy zapisać małymi literami. Opisy artykułów często zawierają dodatkowe formatowanie i ilustracje; przed przystąpieniem do określania podobieństwa należało je usunąć.

Dla wektoryzacji i późniejszego określenia podobieństwa można wykorzystać dowolną bibliotekę z wielu dostępnych dla różnych języków programowania. Przykładem może być SciKit-learn[22] dla języka Python.

Podobieństwa zostały obliczone dla dwóch kolekcji dokumentów. W pierwszej dokumentem był tylko tytuł wpisu; w drugiej – tytuł i opis potraktowane jako całość. W procesie tokenizacji pominięto znaki interpunkcyjne oraz zignorowano wielkość liter. Współczynniki *tf* i *idf* były wygładzane logarytmicznie.

[22] SciKit-learn jest otwartą biblioteką dla Python, dostępną na http://scikit-learn.org. Zawiera implementacje wielu algorytmów stosowanych w tzw. uczeniu się maszyn (ang. *machine learning*) z zakresu klasyfikacji, regresji, klastrowania, redukcji liczby wymiarów i in.



Rys. 8. Wizualizacja podobieństwa kolekcji ok. dwóch tysięcy tytułów doniesień prasowych. Każdy wierzchołek odpowiada jednemu tytułowi. Wielkość i kolor wierzchołka zależą od tego, jak wiele podobnych tytułów znajduje się w kolekcji. Wierzchołki czarnej (najmniejsze) reprezentują tytuły wcale niepodobne do innych (lub podobne w bardzo niewielkim stopniu); pozostają one w głębi grafiki. Bliżej (z perspektywy oglądającego) znajdują się grupy (klastry) wierzchołków czerwonych. Tytuły tworzą klastry o tym większej gęstości („ciaśniej" upakowane grupy), im bardziej są do siebie podobne. Odległość pomiędzy klastrami na tyle, na ile było to możliwe do odwzorowania w trójwymiarowej przestrzeni, odwrotnie proporcjonalna do podobieństwa pomiędzy klastrami. Z czytelniczego punktu widzenia można przypuszczać, że zamiast ok. 14 artykułów (jest to średni rozmiar klastra) w zupełności wystarczyłaby lektura zaledwie jednego (losowo wybranego tzw. reprezentanta grupy) lub np. dwóch będących najbardziej oddalonymi od siebie w ramach grupy (źródło: własne)

5. Wizualizacja kolekcji doniesień prasowych na podstawie wzajemnego podobieństwa

Dla wizualizacji dokumenty i podobieństwa pomiędzy nimi potraktowano jako wierzchołki i nieskierowane ważone krawędzie w grafie spójnym. Dla rozmieszczenia grafu przyjęto, że odległość pomiędzy wierzchołkami ma być tym mniejsza, im większe podobieństwo pomiędzy dokumentami.

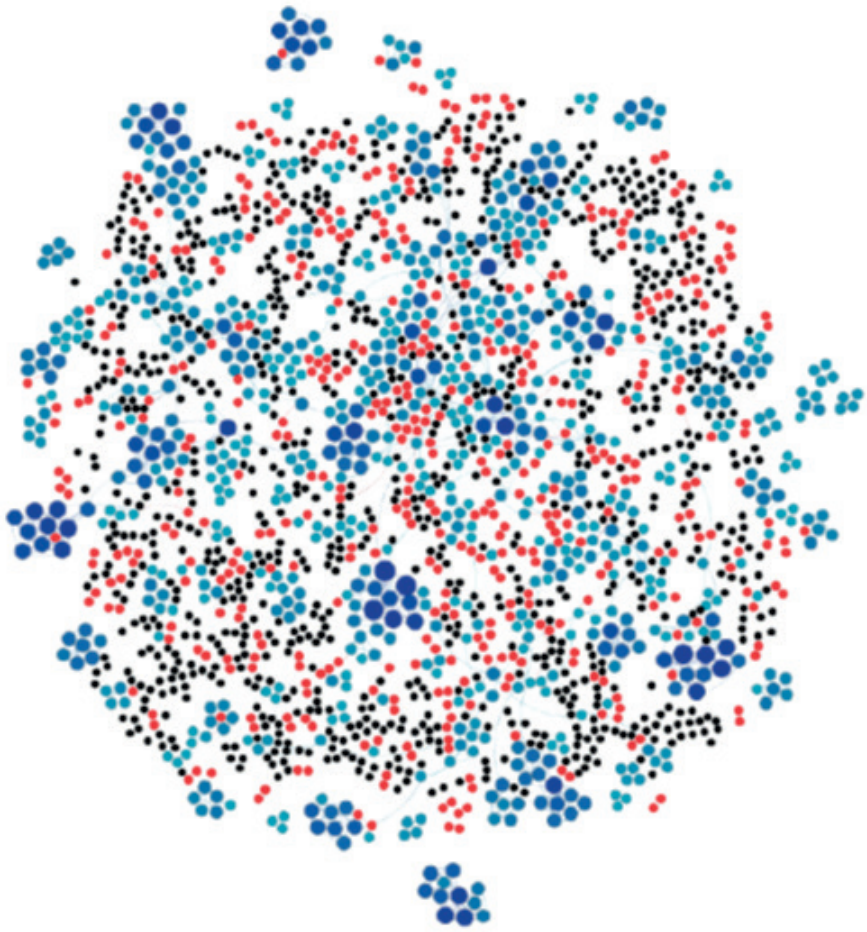
Należy zauważyć, że dostępne jest podobieństwo dla każdej pary dokumentów. Dla otrzymania wizualizacji widocznych na rys. 8 i 9 zredukowano liczbę krawędzi, pomijając te o wagach mniejszych niż 0,2. W wizualizacji otrzymanego grafu (niespójnego podgrafu początkowego grafu pełnego) wielkość wierzchołka zwiększa się wraz z jego stopniem. Dodatkowo stopień wierzchołka identyfikowany jest kolorem, przy czym kolor czarny zarezerwowany jest dla stopnia 0.

Ustanowiono etykietowanie wierzchołków grafu polegające na przypisaniu każdemu z nich „masy" zależnej od wielkości (a więc stopnia) wierzchołka. W wizualizacji podobieństwa tytułów wpisów (rys. 8) skupiska masy umieszczono bliżej obserwatora, co w dwuwymiarowym rzucie oznaczono kolorem czerwonym. Natomiast w wizualizacji podobieństwa tytułów i opisów artykułów (rys. 9), dla zwiększonej czytelności dwuwymiarowego rzutu, kolorem czerwonym oznaczono skupiska bardziej odległe od obserwatora, a bliższe mu – kolorem niebieskim.

6. Interpretacja wizualizacji i uwagi końcowe

W opisanym eksperymencie użyto próbki danych z jednego tygodnia, liczącej ponad 2000 wpisów. Każdy z nich oznacza jeden artykuł na pewnej stronie internetowej, ale nie mówi nic o objętości owego artykułu. Gdyby przyjąć, że każdy z nich ma średnio długość choćby jednej trzeciej długości artykułów, jakie znaleźć można na stronach Państwowej Agencji Prasowej[23], dałoby to łącznie ok. 700 stron maszynopisu.

[23] Artykuły w serwisie http://www.pap.pl mają długość około jednej strony druku w formacie A4.



Rys. 9. Wizualizacja podobieństwa kolekcji ok. dwóch tysięcy tytułów i opisów doniesień prasowych. Porównanie do siebie dokumentów, z których każdy składa się z tytułu i opisu artykułu, skutkuje grafem o większej gęstości, niż można było zaobserwować w przypadku samych tytułów (por. rys. 8). Wyraźnie zauważalna jest większa liczba klastrów (grup skupiających artykuły o podobnej tematyce) niż w przypadku porównania do siebie samych tytułów (por. rys. 8). Każdy wierzchołek odpowiada jednemu artykułowi (dokumentowi), a jego wielkość zależy od liczby podobnych dokumentów w kolekcji. Klastry bliższe z perspektywy obserwatora oznaczono kolorem niebieskim; dalsze – czerwonym. Wierzchołki czarne reprezentują artykuły wcale niepodobne do innych lub podobne w stopniu znikomym. Porównywanie tytułów wraz opisami, w stosunku do porównywania samych tytułów, skutkuje wyodrębnieniem większej liczby klastrów o mniejszej średnicy. Z czytelniczego punktu widzenia oznacza to, że lektura reprezentantów (a więc po jednym artykule z każdej grupy) może oznaczać efektywne zapoznanie się z niemal każdym tematem poruszonym w artykułach znajdujących się w kolekcji (źródło: własne)

Lektura całości tekstu w ciągu tygodnia z pewnością stanowiłaby wyzwanie dla części osób i za uzasadnione można uznać starania minimalizacji tej liczby, przy możliwym zachowaniu wartości informacyjnej.

Wizualne przedstawienie numerycznego porównania informacji o dostępnych artykułach sugeruje, że jest pewna liczba artykułów, które różnią się jedynie nieznacznie i traktują o tych samych wydarzeniach. Przekonują o tym rys. 8 i 9, gdzie dokumenty o znacznym wzajemnym podobieństwie tworzą ciasne skupiska.

Należy zauważyć, że trudno jest precyzyjnie określić skuteczność opisanej metody przez wzgląd na subiektywizm określenia „podobne". Jednak jeśli przyjąć, że teksty są „podobne", gdy traktują o tych samych bytach lub akcjach, to założenie zbieżności słownictwa (przy normalizacji tekstu, a więc ewentualnym uwzględnieniu synonimii, uspójnieniu pisowni np. dat lub liczb, zredukowaniu różnic fleksyjnych itd.) zdaje się zgodne z ludzkim rozumieniem tego pojęcia.

Na rys. 9 przedstawiającym wizualizację podobieństwa tytułów i opisów artykułów wyraźnie zauważalna jest większa liczba klastrów niż w przypadku wzajemnego porównania samych tytułów (rys. 8), co jest to intuicyjnie zrozumiałe, gdyż w dłuższych dokumentach można zaobserwować więcej podobieństw niż w krótkich frazach. Pewnym czynnikiem może być tutaj fakt, że różnice widoczne w „chwytliwych" nagłówkach zanikają już na etapie opisu (streszczenia lub fragmentu) artykułu.

Na każdej wizualizacji część spośród 2000 punktów, w skutek znacznego podobieństwa porównywanych dokumentów, tworzy wyraźnie widoczne grupy. Można przypuszczać, że zamiast losowego próbkowania kolekcji – dla szybszego zapoznania się z nią – znacznie lepiej [pod względem precyzji (ang. *precision*) i pokrycia (ang. *recall*) rozumianych tu w kategoriach *data mining*] byłoby posłużyć się reprezentantami (wybranymi dokumentami – po jednym losowo z każdej grupy). Mając na uwadze obserwowane średnice klastrów, skutkowałoby to oczywiście kilkukrotnym, lub nawet kilkunastokrotnym ograniczeniem liczby tekstów (bez utraty ogólności).

Założenia przyjęte dla zwizualizowania kolekcji są bliskie tym, które przyjmuje się w algorytmach automatycznego grupowania lub klasyfikacji. Istnieje więc możliwość zbudowania narzędzia, które identyfikowałoby grupy podobnych treści programowo (bez wykorzystania strony wizualnej). To pozwoliłoby zbudować narzędzie, które oferowałoby możliwości eksplorowania kolekcji dokumentów „wszerz", a więc przekrojowo, a w przypadku znalezienia interesującego tematu również „w głąb".

Słowa kluczowe: *doniesienia prasowe, pozyskiwanie informacji, deduplikacja informacji.*

Summary

News analysis with natural language processing techniques

The dynamic development of natural language processing results in a growing number of products utilizing so-called speech and language technologies. On the one hand this refers to the possibility of interacting with a computer using a language that people naturally use in speech and writing; one the other – making the information contained in all sorts of texts accessible for a computer.

We present how methods for gathering and extracting information can be applied to news releases, to possibly reduce the overhead generated by republishing the same news by numerous internet information portals.

We present how web syndication can be used to gather press releases; how to process those texts in order to determine mutual similarity; and how to visualize those. We present preliminary results of an experiment with application of the above-mentioned methods to selected Polish internet portals.

Keywords: *press releases, information retrieval, information deduplication*.

Bibliografia

Opracowania

Billsus D., Pazzani M.J, *A Hybrid User Model for News Story Classification,* w: *UM99 User Modeling*. *CISM International Center for Mechanical Sciences*, t. 407, ed. by J. Kay, Wien 1999.

Cooper A., Reimann R., Cronin D., *About Face 3: The Essentials of Interaction Design 3rd Edition*, Indianapolis 2007.

Elbadrawy A., Karypis G., *User-Specific Feature-Based Similarity Models for Top–n Recommendation of New Items*, "ACM Transactions on Intelligent Systems and Technology" 2015, t. 3(6).

Manning Ch., Raghavan P., Schütze H., *An Introduction to Information Retrieval*, Cambridge 2009.

Phuvipadawat S., Murata T., *Breaking News Detection and Tracking in Twitter*, "Web Intelligence and Intelligent Agent Technology" 2010, t. 3.

Internet

http://www.alexa.com/topsites.

http://www.alexa.com/topsites/countries/PL.

https://pypi.python.org/pypi/feedparser.

http://www.similarweb.com/global.

http://o2.pl.

http://www.rssboard.org/rss-0-9-0.

http://www.rssboard.org/rss-0-9-1-netscape.

http://web.resource.org/rss/1.0/.

http://www.rssboard.org/rss-specification.

https://tools.ietf.org/html/rfc4287.
https://tools.ietf.org/html/rfc5023.
http://www.w3.org/TR/2008/REC-xml-20081126/.
http://www.w3.org/TR/2006/REC-xml11-20060816/.
https://www.python.org/.
http://scikit-learn.org.
http://www.pap.pl.