



POLITECHNIKA WARSZAWSKA



WYDZIAŁ MATEMATYKI
I NAUK INFORMACYJNYCH

PRACA DYPLOMOWA MAGISTERSKA
INFORMATYKA

**Rekomendacje artykułów opisujących produkty w
serwisach e-commerce**

Content-based recommendations in e-commerce services

Autor:

Łukasz Dragan

Promotor: dr inż. Anna Wróblewska

Warszawa, czerwiec 2017

.....

podpis promotora

.....

podpis autora

Streszczenie

Tematyka niniejszej pracy skupia się wokół zagadnień określania podobieństwa semantycznego pomiędzy dokumentami tekstowymi i rekomendacji dokumentów podobnych do danego. Szczegółowy problem pochodzi z internetowego serwisu aukcyjnego Allegro, który posiada dział artykułów opisujących produkty dostępne w serwisie. W dziale tym funkcjonuje system rekomendacji podobnych artykułów tekstowych w oparciu o ich treść. Celem pracy jest zbadanie możliwości usprawnienia działania istniejącego systemu rekomendacji.

W niniejszej pracy adaptuję dostępne metody określania podobieństwa pomiędzy dokumentami tekstowymi do powyższego problemu, wprowadzam miary umożliwiające ocenę działania tych metod oraz dokonuję analizy możliwości ich wykorzystania w rzeczywistym systemie. Skupiam się szczególnie na metodach opartych o semantyczną analizę tekstu.

Abstract

Spis treści

1	Wstęp	4
1.1	Rekomendacje artykułów tekstowych w Allegro	5
1.2	Struktura pracy	7
1.3	Uwagi	7
1.3.1	7
2	Przegląd wybranych metod	8
2.1	Systemy rekomendacji	8
2.1.1	Filtrowanie kolaboratywne (collaborative filtering)	9
2.1.2	Filtrowanie oparte na treści (content-based filtering)	9
2.2	Techniki przetwarzania języka naturalnego	10
2.2.1	Bag-of-words	10
2.2.2	Term frequency - inverted document frequency	11
2.2.3	Latent Dirichlet Allocation	12
2.2.4	Word2vec	12
2.2.5	Odległość między dokumentami	13
3	Dane	16
3.1	Wstępne przetwarzanie danych	16
3.2	Opis danych po wstępnym przetwarzaniu	18
4	Metody ewaluacji	20
4.1	Miara 1: Dystans oparty na metadanych	21
4.1.1	Kategorie	21

4.1.2	Słowa kluczowe	23
4.2	Miara 2: Ocena przez użytkowników offline	23
4.3	Miara 3: Historyczna aktywność użytkowników serwisu	23
5	Opis i wyniki badań	26
5.0.1	26
5.1	Dalsze badania	26
A	Technologie i narzędzie	28

Rozdział 1

Wstęp

Systemy rekomendacji są powszechnym elementem wielu serwisów internetowych. Sprawdzają się na takich polach, jak polecanie produktów w sklepie czy rekomendacje ofert pracy. Dają użytkownikowi poczucie indywidualnego traktowania przez serwis internetowy dopasowujący niejako zawartość swoich stron to konkretnego użytkownika. Pozwala to użytkownikowi na bardziej efektywne korzystanie z serwisu. Może to prowadzić do większego zaangażowania ze strony użytkownika i przywiązania do serwisu. Systemy rekomendacji dają obopólną korzyść zarówno użytkownikowi jak i właścicielowi serwisu internetowego.

Celem niniejszej pracy magisterskiej jest analiza możliwości usprawnienia istniejącego systemu rekomendacji o oparciu o adaptację istniejących metod wyszukiwania semantycznego podobieństwa pomiędzy dokumentami tekstowymi. Rzeczony system rekomendacji istnieje w internetowym serwisie e-commerce Allegro w dziale artykułów tekstowych o tematyce związanej z produktami dostępnymi za pośrednictwem serwisu. System ma na celu zarekomendowanie użytkownikowi artykułów o tematyce podobnej do tego, który znajduje się na stronie aktualnie odwiedzanej przez użytkownika.

W swojej pracy badam możliwość użycia istniejących metod semantycznej analizy tekstu w odniesieniu do opisanego problemu. Badane metody to Latent Dirichlet Allocation oraz word2vec z wariantem Word Mover's Distance oraz bez

niego.

Podczas prowadzenia badań stworzyłem szereg skryptów przetwarzających dane i wykorzystujących implementacje opisywanych w tej pracy metod. Opis użytych narzędzi programistycznych i bibliotek zawarłem w dodatku A do niniejszej pracy.

1.1 Rekomendacje artykułów tekstowych w Allegro

Allegro jest największą działającą na rynku polskim platformą aukcyjną on-line. Posiada ponad 20 mln zarejestrowanych klientów. Każdego dnia na Allegro sprzedaje się ponad 870 tysięcy przedmiotów. Zatrudnia 1300 pracowników.[3] Serwis umożliwia użytkownikom wystawianie na sprzedaż oraz kupno przedmiotów poprzez mechanizm licytacji lub natychmiastowego zakupu. Allegro pobiera prowizję za dokonanie sprzedaży za swoim pośrednictwem.

Oprócz głównej części serwisu odpowiedzialnej za transakcje Allegro posiada dział zajmujący się publikacją artykułów opisujących produkty wystawiane za pośrednictwem serwisu. Ma to na celu pomoc użytkownikom przy wyborze interesującego ich produktu.

Po to, aby zachęcić użytkowników do zapoznania się z treścią kolejnych artykułów, zastosowany został tu system rekomendacji przyporządkowujący danemu artykułowi listę powiązanych artykułów. Kryterium mówiącym, czy artykuły są powiązane jest tutaj jedynie treść artykułów a nie wcześniejsze zachowanie użytkownika.

The screenshot shows the Allegro website interface. At the top, there's a navigation bar with links like 'Strefa Marek', 'Inspiracje', 'Poradniki', 'wystaw przedmiot', 'obserwowane', 'moje allegro', 'załóż konto', and 'zaloguj'. Below this is the Allegro logo and a search bar with the text 'czego szukasz?'. A row of category links follows: 'Elektronika', 'Moda i uroda', 'Dom i zdrowie', 'Dziecko', 'Kultura i rozrywka', 'Sport i wypoczynek', 'Motoryzacja', 'Kolekcje i sztuka', 'Firma', and 'Strefa okazji'. The main content area features an article titled 'Jaka farba dla alergika?' with a sub-header 'Allegro - Poradniki - Dom i zdrowie - Jaka farba dla alergika?'. The article includes a large image of a room with four colored squares (blue, purple, yellow, red) on a white wall and a green ladder. The author's name, Ewelina Wojtunik, is displayed with a small profile picture and a bio. The article text discusses the health benefits of low-VOC paints for allergy sufferers. A sidebar on the right contains a section titled 'może Cię również zainteresować' (You might also be interested in) with a link to 'Wnętrzarski hit – ściany ombre' (Interior hit – ombre walls).

Jaka farba dla alergika?

autor: Ewelina Wojtunik, data publikacji: 23-04-2015

Za chwilę wiosna, a wraz z nią potrzeba porządków i odświeżenia ścian. Jak co roku będziemy sprzątać, wietrzyć i wymiatać zimę z kątów mieszkania. Zaraz po tym zaczną się pierwsze remonty.

Czy wiecie, że nawet do 30% populacji to alergicy? Uczulają nas pyłki, kurz, sierść zwierząt, pokarmy, ale przede wszystkim chemikalia i detergenty. Znajdujące się w nich alergeny mogą być powodem problemów zdrowotnych, a także nasilać objawy nadwrażliwości takie jak łzawienie oczu, zapalenie skóry czy kaszel astmatyczny.

Szkodliwe związki lotne

W styczniu 2010 roku Unia Europejska wprowadziła normę, która reguluje zawartość szkodliwych lotnych związków organicznych tak zwanych LZO (VOCs, ang. volatile organic compounds) w trafiających do sprzedaży farbach i lakierach. Warto wiedzieć, że lotne związki lubią pozostawiać aktywne pomimo wyschnięcia farby i starannego wywietrzenia mieszkania. Co więcej, mogą uwalniać się ze ścian całymi latami, nasilając objawy alergiczne i pogarszając samopoczucie mieszkańców. Im mniej ich w składzie, tym lepiej dla nas.

może Cię również zainteresować

Wnętrzarski hit – ściany ombre

Ombre stało się hitem w wizażu i modzie już kilka sezonów temu! Chętnie rozjaśniamy końcówki włosów, cieniujemy kolory na paznokciach, a także nosimy ubrania w przenikających si...

Rysunek 1.1: Widok strony internetowej zawierającej jeden z artykułów serwisu Allegro. [14]

Od serwisu Allegro otrzymałem zserializowaną kopię 20000 artykułów dostępnych na stronach serwisu. Pojedynczy artykuł składa się z głównej zawartości tekstowej oraz metadanych. W celu otrzymania wszelkich danych od firmy Allegro wynagane było, abym podpisał umowę, w której zobowiązuje się do nieujawniania żadnych danych, które otrzymałem. Stąd opisy danych, na których pracuję, zawarte w tej pracy nie wnikają w ich szczegóły i nieodciągają od informacji publicznie

dostępnych za pośrednictwem strony pod adresem `/urlhttps://allegro.pl`.

1.2 Struktura pracy

Rozdział 2 wprowadza do zagadnienia rekomendacji i opisuje wybrane metody przetwarzania języka naturalnego służące do wyszukiwania podobieństw między dokumentami tekstowymi.

Następnie w rozdziale 3 dokonuję opisu konkretnego problemu, jakim jest generacja rekomendacji artykułów tekstowych w serwisie Allegro. Opisuję dane otrzymane z serwisu oraz kolejne wstępne etapy przetwarzania ich, aby nadawały się do zaaplikowania do nich wybranych metod.

Dalej, w rozdziale 4 opisuję stworzone i zastosowane później metody ewaluacji wyników.

Następnie w rozdziale 5 opisuję proces zastosowania wybranych metod oraz porównanie efektów ich działania.

Ostatecznie w rozdziale 6 dokonuję podsumowania przeprowadzonych badań i rozważam kierunki dalszych prac w tej dziedzinie.

Załącznik A zawiera opis narzędzi programistycznych i bibliotek wykorzystanych przez mnie podczas prowadzenia badań.

1.3 Uwagi

1.3.1

W celu uniknięcia nieporozumień pragnę tutaj zaznaczyć różnicę pomiędzy znaczeniami słowa „artykuł”, które może oznaczać zarówno tekst publicystyczny, literacki lub naukowy jak i rzecz, która jest przedmiotem handlu.[2] W niniejszej pracy skupiam się na rekomendacjach artykułów tekstowych, stąd używam pierwszego znaczenia (chyba, że inne znaczenie jest wyraźnie zaznaczone).

Rozdział 2

Przegląd wybranych metod

W swojej pracy wykorzystuję i adaptuję do swoich potrzeb szereg metod i narzędzi umożliwiających przetwarzanie języka naturalnego, semantyczną analizę tekstu i wykrywanie podobieństwa pomiędzy tekstami. Część z nich (metoda tf-idf, bag-of-words, silnik Elasticsearch) jest od lat powszechnie wykorzystywana w zadaniu wyszukiwania tekstowego. Inne z kolei - korzystające z semantycznej analizy tekstu - nie są tak popularne z powodu swojej nowości, bądź trudności w zaaplikowaniu. Daje to pole do badań i ewentualnych usprawnień istniejących systemów opierających się o klasyczne metody. Wybrane metody stosuję, zgodnie z tematem pracy, w zadaniu generowania rekomendacji, stąd przegląd metod zaczynam właśnie od wprowadzenia do tego zagadnienia.

2.1 Systemy rekomendacji

Systemy rekomendacji to narzędzia i techniki mające na celu zasugerować użytkownikowi przedmioty. Sugestie te odnoszą się do różnych procesów podejmowania decyzji takich jak np. które artykuły kupić, jakiej muzyki słuchać czy też które wiadomości czytać. „Przedmiot” jest tutaj ogólnym pojęciem oznaczającym coś, co system poleca użytkownikowi. [1]

Przy wciąż wzrastającej ilości danych użytkownicy serwisów internetowych często nie są w stanie dotrzeć do informacji, która ich interesuje. Jest to

pole do rozwoju zautomatyzowanych systemów rekomendacyjnych polecających użytkownikom treści, które mogą ich zainteresować. Działalność takiego systemu daje zysk zarówno użytkownikowi, pozwalając mu dotrzeć do informacji, której mógłby samodzielnie nie odszukać, albo wręcz nie wiedzieć, iż taka informacja istnieje, jak i dla właścicieli serwisów internetowych, którym zależy, by przyciągnąć do siebie użytkowników, aby ci w jak największym stopniu korzystali z ich usług.

Sposoby działania systemów rekomendacji można podzielić na różne sposoby, spośród których wyodrębnić można dwa najszerzej używane. Są to: filtrowanie kolaboratywne (collaborative filtering) i filtrowanie oparte na treści (content-based filtering).

2.1.1 Filtrowanie kolaboratywne (collaborative filtering)

Technika ta opiera się na spostrzeżeniu, iż użytkownicy o podobnych preferencjach zachowują się podobnie. Stąd jeżeli użytkownik zachowuje się podobnie do zaobserwowanej wcześniej grupy użytkowników, można przewidzieć jego preferencje. Istotną zaletą tej metody jest fakt, iż nie zależy ona od dziedziny, w której ulokowany jest system rekomendacji (w przeciwieństwie do rekomendacji opartych na treści), a jedynie od zachowań użytkowników.

2.1.2 Filtrowanie oparte na treści (content-based filtering)

W technice tej przedmioty polecane użytkownikowi zależą od innych przedmiotów, na temat których stwierdzono, że użytkownik się nimi interesuje. Mogą się one opierać np. na podobieństwie przedmiotów: jeżeli użytkownik „lubi” przedmiot A, który jest podobny do przedmiotu „B” to można spodziewać się, że również przedmiot B zainteresuje użytkownika. Technika ta jest mocno zależna od dziedziny rekomendowanych przedmiotów, gdyż wymaga wprowadzenia pewnej miary podobieństwa między nimi. Stąd jest trudniejsza do zastosowania, ale daje też możliwości nieosiągalne dla filtrowania kolaboratywnego.

Celem niniejszej pracy jest zbadanie metod sugerujących użytkownikowi

artykuły podobne do aktualnie odwiedzanego, co wprost wiąże się z metodami używanymi w technice filtrowania opartego na treści.

2.2 Techniki przetwarzania języka naturalnego

Temat niniejszej pracy skupia się na podobieństwie pomiędzy artykułami - dokumentami tekstowymi. Ich treść zapisana jest w języku naturalnym - zrozumiałym dla człowieka - który mówiąc potocznie niezrozumiały dla maszyny. W związku z tym koniecznym staje się tu użycie technik przetwarzania języka naturalnego (natural language processing), które to pozwalają wyodrębnić z tekstu pewne cechy, na bazie których komputer jest w stanie określić podobieństwo pomiędzy dokumentami (według pewnej sformalizowanej miary). W poniższych paragrafach opisuję techniki przetwarzania języka naturalnego użyte przeze mnie wprost lub z których koncepcji czerpałem.

W celu formalizacji dalszych opisach stosowanych metod stosuję następujące oznaczenia:

- Korpus C : zbiór dokumentów d ,
- Dokument d : skończony ciąg zdań s ,
- Zdanie s : skończony ciąg słów w ,
- Słowo w : skończony ciąg znaków c ,
- W celu uproszczenia zapisu: $w \in d \equiv \exists_{s \in d} w \in s$,
- Słownik zbudowany na korpusie C : $V = w \mid \exists_{d \in C} w \in d$.

2.2.1 Bag-of-words

Bag-of-words (worek słów) jest metodą reprezentacji tekstu jako zbioru zawartych w nim słów niezachowującego kolejności słów w tekście, lecz liczbę

ich wystąpień. Jako korpus będę nazywać zbiór przetwarzanych dokumentów, natomiast jako słownik zbiór słów

Bag-of-words można opisać jako przekształcenie z korpusu w przestrzeń wektorów $bow : C \rightarrow \mathbb{R}^n$ gdzie:

C : korpus

$m = |C|$: liczba dokumentów w korpusie C

V : słownik zbudowany na C

$n = |V|$: liczba słów w V

$v_i \in \mathbb{R}^n$, gdzie $i \in 1, 2, \dots, n$ wektor reprezentujący dokument $d_i \in C$

v_{ij} , gdzie $j \in 1, 2, \dots, m$: liczba wystąpień w dokumencie $d_i \in C$ słowa $w_j \in V$

Technika ta jest stosunkowo prosta jest jej wadą jest traktowanie każdego słowa z jednakową wagą. Pewne słowa (np. „i”, „lub”, „o”) występują bardzo często, lecz ich wkład w znaczenie całego dokumentu jest marginalny. Stąd powstały bardziej zaawansowane techniki uwzględniające istotność słów dla znaczenia całego dokumentu.

2.2.2 Term frequency - inverted document frequency

TF-IDF (ważenie częstością termów - odwrotna częstość w dokumentach) jest metodą reprezentacji tekstu jako zbioru słów przy jednoczesnym uwzględnieniu wagi słów, która zależy od częstości występowania słowa w korpusie. Oznaczenia formalne takie same tak w przypadku BOW. $v_{ij} = tfidf_{ij} = tf_{ij} * idf_i$, gdzie:

$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$, „term frequency” to liczba wystąpień słowa w_i w dokumencie d_j podzielona przez liczbę słów dokumentu d_j ,

$idf_i = \log \frac{|D|}{|d: w_i \in d|}$, „inversed document frequency” to liczba dokumentów w korpusie podzielona przez liczbę dokumentów zawierających przynajmniej jedno wystąpienie słowa w_i ,

W tej technice słowa występujące rzadko są premiowane względem słów pospolitych.

2.2.3 Latent Dirichlet Allocation

2.2.4 Word2vec

Word2vec jest stosunkowo nową (2013) metodą osadzania słów w przestrzeni wektorowej (word embedding), opisaną w [5].

Autorzy metody proponują płytką, dwuwarstwową sieć neuronową, która ma za zadanie odtworzyć kontekst danego słowa. Jako wejście metoda otrzymuje słowa z korpusu. Wyjściem metody są natomiast wektory z pewnej N wymiarowej przestrzeni odpowiadające słowom składającej się z warstw: wejściowej, jednej warstwy ukrytej i warstwy wyjściowej. Wyróżnia się dwie architektury sieci: skip-gram: na podstawie słowa sieć przewiduje N sąsiednich słów lub CBOW: na podstawie okna N sąsiednich słów sieć przewiduje słowo, którego z największym prawdopodobieństwem te N słów jest sąsiedztwem. Wady i zalety obu podejść są wymienione w [6]

Softmax jest generalizacją funkcji logistycznej, zamieniającą K -wymiarowy wektor z dowolnych liczb rzeczywistych na K -wymiarowy wektor liczb rzeczywistych z zakresu $(0, 1]$, które sumują się do 1 [13]. Funkcja wyraża się wzorem $\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}$ dla $j = 1, \dots, K$. Wyjście funkcji można traktować jako pewien rozkład prawdopodobieństwa.

Używając tej stosunkowo prostej architektury można wykonać proces nauki używając milionów słów, których powiązania między sobą zostaną zachowane w systemie wag sieci neuronowej.

W metodzie word2vec nauka polega na trzelenowaniu sieci neuronowej. Jednakże w odróżnieniu od innych metod wykorzystujących sieci neuronowe, word2vec nie używa później wytrenowanej sieci jako takiej, a jedynie otrzymanych w wyniku nauki wag warstwy ukrytej sieci, które faktycznie są wynikowymi wektorami słów.

W dalszym opisie metody szczegółowo skupiam się na podejściu CBOW, lecz podejście skip-gram wygląda analogicznie.

Sieć neuronowa będąca wynikiem nauki przyjmuje na wejściu wektor binarny długości odpowiadającej liczbie słów w słowniku V zbudowanym na korpusie

treningowym. Wektor ten wypełniony jest wartościami 0 oraz jedną wartością 1 na i -tej pozycji. Taki wektor odpowiada i -temu słowu ze słownika V . Wejściem sieci są kolejne słowa z korpusu. Wyjściem sieci jest wektor tej samej długości o wartościach rzeczywistych z zakresu $[0,1]$, w którym wartość na i -tej pozycji odpowiada prawdopodobieństwu, że i -te słowo ze słownika znajduje się w sąsiedztwie słowa wejściowego. Za „sąsiedztwo” wielkości x należy tu rozumieć zbiór złożony z x słów występujących przed danym słowem w korpusie i x słów położonych za danym słowem. Wartość x może być tu ograniczona przez początek/koniec zdania, które ograniczają kontekst danego słowa.

Jako efekt należy się spodziewać, że dla słowa wejściowego „Brytania” otrzymamy na wyjściu wysoką wartość prawdopodobieństwa dla słowa „Wielka”, a niską np. dla słowa „skoroszyt”.

Jednym z parametrów metody word2vec jest wymiarowość przestrzeni, w której znajdują się otrzymane wektory odpowiadające słowom z korpusu. Liczba ta ma swoje źródło z wielkości warstwy ukrytej sieci neuronowej. Wagi warstwy ukrytej można interpretować jako macierz $M \times N$, gdzie M to liczba słów słownika V - wielkość wektowa wejściowego, a N to liczba neuronów w warstwie ukrytej. Po przeprowadzeniu nauki i -ty wiersz tej macierzy odpowiada wektorowi długości N , który reprezentuje i -te słowo ze słownika V .

W sieci nie jest używana funkcja kakywacji, ale prawdopodobieństwa na wyjściu są efektem działania funkcji softmax.

Funkcja softmax ma tutaj za zadanie sprowadzić wyjściowe wartości warstwy ukrytej do postaci rozkładu prawdopodobieństwa.

Użycie metody Word2vec pozwala ocenić „odległość” pomiędzy dwoma dokumentami nawet, jeżeli nie posiadają one wspólnych słów. Jest to metoda osadzania (embedding) słów w pewnej przestrzeni wektorowej.

2.2.5 Odległość między dokumentami

W celu wykorzystania omówionej metody Word2vec w obszarze tematyki pracy należy wybrać metodę obliczania odległości między dokumentami. Zakładamy, że

jeżeli dystans pomiędzy dokumentami jest mały, to ich tematyka jest podobna.

Centroid

Najprostszą i najbardziej intuicyjną metodą obliczenia odległości pomiędzy wektorową reprezentacją dokumentów jest wykonanie dwóch prostych kroków:

1. Uśrednienie wektorów wchodzących w skład każdego z dokumentów. Powstały w ten sposób wektor jest centroidem reprezentującym dokument w przestrzeni wektorowej.

2. Obliczenie dystansu między wektorami. Powszechnie przyjętą praktyką jest stosowanie tzw. odległości kosinusowej - znormalizowanego iloczynu skalarnego wektorów A i B . Jest to kosinus kąta pomiędzy dwoma wektorami reprezentującymi dokumenty. Zaletą tej metody jest natychmiastowa

normalizacja wyniku do zakresu $(0, 1)$. $sim = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$, gdzie A_i i B_i są składowymi wektorów odpowiednio A i B

Wadą opisaną powyżej metody jest utrata potencjalnie użytecznych zależności wektorami wchodzącymi w skład dokumentu.

W kontrze to tego prezentujemy metodę liczenia szukanego dystansu uwzględniającą rozkład wektorów wewnątrz dokumentu.

Word Mover's Distance

Word Mover's Distance[12] to stosunkowo nowe rozwiązanie (2013) zwracające odległość między dokumentami tekstowymi. W tym celu adaptuje algorytm Earth Mover's Distance[10] oraz wektorową reprezentację słów dokumentu. WMD mierzy odległość między dokumentami jako minimalny dystans jaki wektory słów pierwszego dokumentu muszą „pokonać” aby osiągnąć wartości wektorów z drugiego dokumentu.

EMD jest metodą mierzenia odległości pomiędzy dwoma rozkładami, która opiera się na minimalnym koszcie, jaki musi zostać poniesiony, aby dokonać

transformacji jednego rozkładu w drugi. Problem można sformalizować jako problem programowania liniowego, gdzie: $P = \{f(p_1, w_{p_1}) \dots (p_m, w_{p_m})\}$, $Q = \{f(q_1, w_{q_1}) \dots (q_n, w_{q_n})\}$ są danymi rozkładami o m (odpowiednio n) klastrach p_i (q_j), a w_{p_i} (w_{q_j}) jest masą klastra. $D = [d_{ij}]$ jest macierzą odległości, w której d_{ij} reprezentuje odległość pomiędzy klastrami p_i i q_j . Celem jest znaleźć taki przepływ $F = [f_{ij}]$, gdzie f_{ij} to przepływ pomiędzy p_i i q_j , który minimalizuje całościowy koszt $Work(P, Q, F) = \sum_{i=1}^m \sum_{j=1}^n d_{ij} f_{ij}$ przy odpowiednich ograniczeniach[11]. EMD jest to dobrze zbadanym problem transportowym[10], dla którego powstały efektywne metody rozwiązania[9].

Przypuśćmy, że dzięki metodzie word2vec dla słownika V o n słowach otrzymujemy macierz $X \in \mathbb{R}^{d \times n}$. i -ta kolumna tej macierzy reprezentuje i -te słowo ze słownika V . Odległości pomiędzy wektorami reprezentującymi semantycznie zbliżone słowa są relatywnie mniejsze od odległości dla słów niezwiązanych ze sobą. Celem WMD jest zawrzeć semantyczne podobieństwo pomiędzy poszczególnymi parami słów w dystans pomiędzy całymi dokumentami. Aby to osiągnąć metoda traktuje dokument jako rozkład, którego i -tym elementem jest liczba wystąpień i -tego słowa w tym dokumencie, a następnie stosuje metodę EMD do obliczenia dystansu między tymi rozkładami. Macierz odległości D używana w metodzie EMD jest zbudowana na bazie odległości między wektorami Word2vec reprezentującymi słowa dokumentów. $d_{ij} = \|x_i - x_j\|$, gdzie i i j to indeksy słów ze słownika V a x_{ij} to element macierzy X [12]. Autorzy metody określają złożoność metody jako $O(p^3 \log p)$, gdzie p to wielkość słownika V .

Rozdział 3

Dane

Dane, na których testowane były opisywane w niniejszej pracy metody otrzymałem dzięki życzliwości serwisu e-commerce Allegro. Jednak, by dane te otrzymać, zobowiązany zostałem po podpisaniu umowy o poufności. Stąd, w niniejszej pracy brak jakichkolwiek przykładów danych, a jedynie opisy metod użytych do ich przetwarzania i generowania rekomendacji.

Jako, że artykuły ze zbioru dotyczą produktów sprzedawanych za pośrednictwem serwisu Allegro, w skład słownika wchodzi wiele słów specyficznych dla danej branży. Są to m.in. nazwy modeli aparatów:, samochodów:, gier komputerowych:, a także nazwy techniczne: sprężarka,, W związku z tym zachodzi podejrzenie, że zastosowanie metod wykorzystujących model nauczony na ogólnym zbiorze tekstu może nie dawać satysfakcjonujących rezultatów.

3.1 Wstępne przetwarzanie danych

W celu zwiększenia skuteczności metod analizy tekstu stosuje się wstępne przetwarzanie danych. Jego techniki nie wchodzi w skład żadnego standardu. Wykonuje pewne techniki, opisane niżej, zgodnie z intuicją.

Surowe artykuły otrzymane od Allegro posiadają w swej treści wiele znaczników interpretowanych przez system, na podstawie których wzbogacana

jest warstwa wizualna strony internetowej zawierającej artykuł. Np. obrazki czy łącza do ofert związanych z tematem artykułu. Z punktu widzenia semantycznej analizy tekstu są one bezużyteczne, czy wręcz szkodliwe (powodują pewne „zanieczyszczenie” tekstu). Stąd usuwam je wykorzystując odpowiednio skonstruowane wyrażenia regularne (ich postać jest szczegółem nieistotnym z punktu widzenia tematyki niniejszej pracy).

Kolejnym elementem wstępnego przetwarzania tekstu jest usunięcie tzw. słów stopu (ang. stopwords) - na ogół krótkich słów nie wnoszących nic do znaczenia całości artykułu. Są to np. „w”, „z”, „ponieważ”. Ich usunięcie zmniejsza liczbę słów dokumentu skracając tym samym czas jego przetwarzania. Jako że słowa te występują często, usunięcie ich daje możliwość uwypuklenia znaczenia innych słów mających wpływ na rzeczywiste znaczenie całego artykułu.

Następnie sprowadzam wszystkie słowa dokumentu do małych liter, żeby ujednolicić postać części słów o tym samym znaczeniu, wśród których jedno występuje na początku zdania a inne w środku.

Kolejnym, najistotniejszym etapem wstępnego przetwarzania danych jest tzw. tokenizacja, czyli sprowadzanie słów o tym samym znaczeniu, a różnej formie gramatycznej do tej samej postaci. Sporym utrudnieniem jest tutaj stopień skomplikowania języka polskiego oraz liczba wyjątków, jaką ten język posiada. Za przykład może posłużyć słowo „mieć”, którego jedna z form to „ma”, kolejna to „miej”. Celem etapu jest sprowadzenie każdego z tych wyrazów do formy podstawowej „mieć”. Do przeprowadzenia tej operacji stosuję narzędzie Morfologik[4].

Użycie wymienionych technik nie jest jedynym standardem a wynikiem analizy przetwarzanych danych i techniki te zostały dobrane dla konkretnego przypadku

rozbijanie słów połączonych myślnikiem

Po powyższych etapach słownik zbudowany na korpusie zawiera 98174 słów.

3.2 Opis danych po wstępnym przetwarzaniu

Opisać dokładnie pola jsona Napisać o konieczności oczyszczenia tekstu z [werew]

W skład faktycznej treści artykułu wchodzi trzy pola odpowiadające za: zawartość, tytuł i nagłówek. Pozostałe pola wykorzystywane przez mnie pola to: słowa kluczowe i lista kategorii.

Trudności wynikające z przetwarzania języka polskiego

Liczba słów w korpusie Słowa rzadkie itp Rzeczy, które pomijam można zaznaczyć, że są tematem osobnych badań

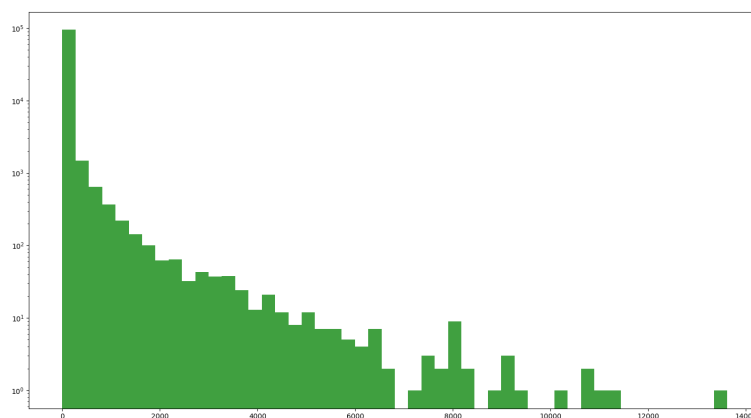
Ewaluacja rankingów jest zadaniem trudniejszym od oceny np. klasyfikatora.

Jaka byłaby sytuacja idealna - w której ocena nie byłaby problemem

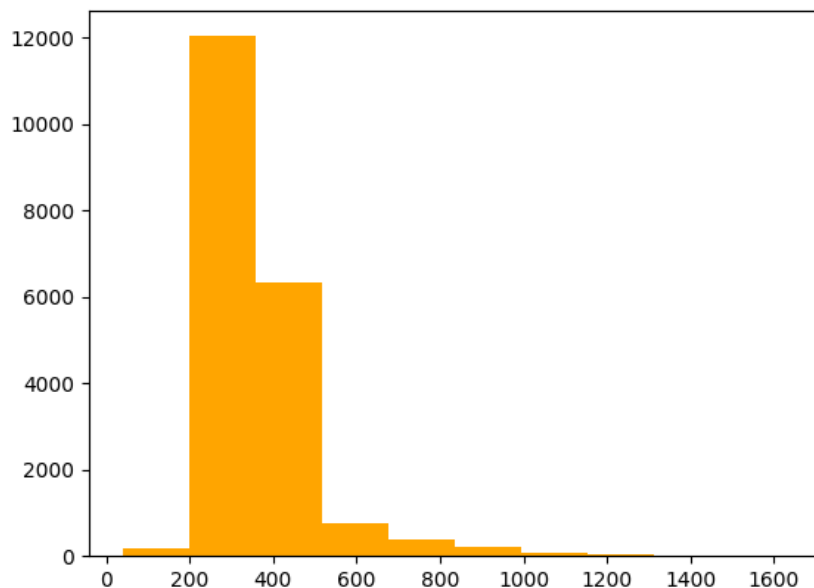
Wspomnieć, że kategorie są drzewiaste

Każdemu artykulowi przypisana jest lista kategorii (zawierających się w sobie pod kątem szczegółowości) klasyfikujących artykuł pod kątem poruszanej tematyki. Wszystkie kategorie tworzą strukturę drzewiastą. Jest to ważny element danych ponieważ pozwala w późniejszym etapie na dokonanie ewaluacji rozwiązania.

Jakość danych: czy nie ma luk Jakość danych oceniam na wysoką, tj. każde pole zawarte w strukturze dokumentu jest zawsze wypełnione - brak jest wartości NULL.



Rysunek 3.1: Histogram wystąpień słów w korpusie



Rysunek 3.2: Histogram długości artykułów.

Otrzymane przeze mnie dane to nieco ponad 20000 dokumentów zapisanych w formacie JSON zawierających główną zawartość artykułu oraz metadane, m.in: id, słowa kluczowe, kategoria, id autora, tytuł, nagłówek.

Rozdział 4

Metody ewaluacji

W celu porównania stosowanych metod wyznaczania podobieństwa między artykułami konieczna jest formalizacja pewnych miar tego podobieństwa.

Ewaluacja rankingu, którym trafność wyników jest zadaniem nietrywialnym. Podobieństwo artykułów napisanych w języku naturalnym jest rzeczą subiektywną. W sytuacji idealnej dysponowalibyśmy obiektywną miarą podobieństwa pomiędzy N artykułami (np. wyznaczoną wcześniej przez miarodajną grupę użytkowników), które to N artykułów stanowiłoby zbiór testowy. Uzyskanie takich danych wiąże się jednak z dużymi kosztami i leży poza możliwościami autora.

Praktyką umożliwiającą obiektywną ocenę, wykorzystywaną w działających systemach są tzw. testy A/B polegające na podziale użytkowników na grupy i zaaplikowaniu każdej grupie innego rozwiązania. Następnie mierzone są pewne wskaźniki wśród każdej grupy (w naszym przypadku np. liczba kliknięć w artykuły rekomendowane) i spośród zgromadzonych wyników wybierane jest rozwiązanie najlepsze.

Z powodu braku możliwości wykorzystania rzeczywistych użytkowników do ewaluacji rozwiązań jestem zmuszony wprowadzić własne miary oparte na dostępnych danych.

Należy tu zaznaczyć niedoskonałość wprowadzanych miar, ponieważ każda z nich opiera się na pewnych założeniach, od których prawdziwości zależy jakość samej miary.

Działanie testowanych metod można sformalizować w postaci pewnej funkcji $S_n : C \rightarrow \{a_i\}_{i < n}$, gdzie $a_i \in C$, a n to liczba elementów zwracanego ciągu. Funkcja S przyjmuje artykuł tekstowy (bądź jego identyfikator) i zwraca skończony ciąg artykułów do niego podobnych zgodnie ze stopniem dopasowania (najlepsze na początku). Celem działania niżej opisanych miar jest każdej parze: wyjście-wejście funkcji S reprezentującej testowaną metodę przypisać ocenę jakości zwróconego wyjścia dla danego wejścia. Ocenę dla konkretnej metody, dla ustalonej próby artykułów są następnie uśredniane.

Opisane poniżej miary 1 i 2 dokonują porównania podobieństwa dla pary artykułów. W celu rozszerzenia działania tych miar do pary wejście-wyjście metody stosuje średnią ważoną podobieństwa kolejnych elementów wyjścia z wejściem. Stosowane wagi: $\frac{1}{i}$ dla $i = 1, \dots, N$, gdzie N to długość ciągu wyjściowego danej metody.

4.1 Miara 1: Dystans oparty na metadanych

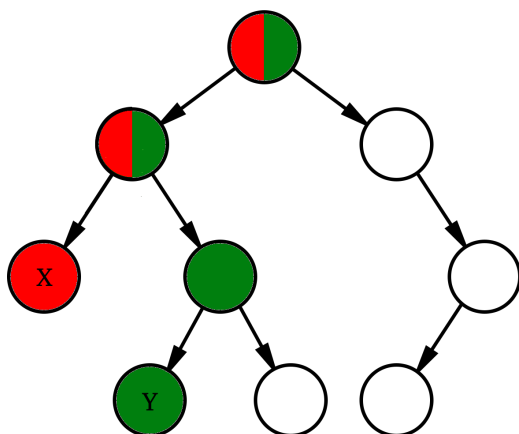
Jak wspomniałem wcześniej dane prócz treści artykułów zawierają również pewne metadane, a wśród nich umożliwiające tworzenie powiązań między artykułami. Skupię się tu na dwóch polach: „słowa kluczowe” i „kategoria”.

4.1.1 Kategorie

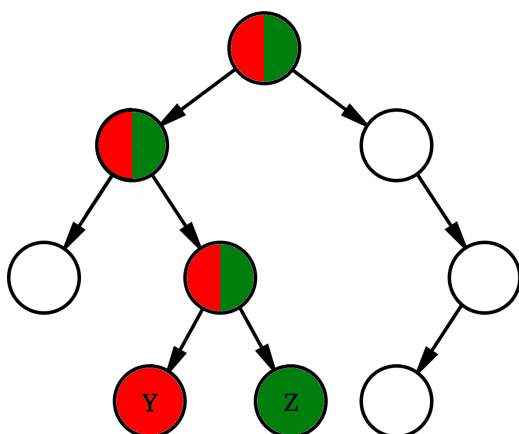
Pierwszą zastosowaną miarą, pozwalającą ocenić jakość dopasowania podobnych artykułów jest ich odległość we wcześniej wspomnianym drzewie kategorii. Zakładam tu, że im więcej wspólnych przodków w drzewie oraz im mniej przodków rozbierzmych, tym bardziej podobne do siebie są artykuły reprezentowane przez węzły drzewa. Zaletą miary jest fakt, iż przypisanie artykułu do kategorii zostało wykonane przez autora, którego można określić ekspertem w danej dziedzinie. Stąd przynależność artykułu do danej kategorii jest mocno uzasadniona. Kolejną zaletą tej miary jest fakt, iż można ją zastosować automatycznie - wiedza ekspercka jest już zapisana w danych artykułów. Należy

zaznaczyć tu jednak, że miara nie jest idealna - każdy artykuł należy do tylko jednego liścia drzewa kategorii. Stąd artykuł poruszający zagadnienia z różnych obszarów, który można by przypisać dwóm stosunkowo odległym kategoriom A i B , zostanie przypisany tylko do jednej kategorii, np. A . Miara pokaże wtedy dużą odległość od artykułów z kategorii B , co nie jest prawdą.

Formalnie miarę można zapisać jako: $d(a_1, a_2) = \frac{w(a_1, a_2)}{D}$, gdzie d to dystans między artykułami a_1 i a_2 , $w(x, y)$ to długość części wspólnej ścieżek od korzenia drzewa kategorii do węzłów reprezentujących artykuły x i y , a D to głębokość całego drzewa (wprowadzone w celu normalizacji). Im wyższy wynik, tym większe podobieństwo artykułów.



Rysunek 4.1: Drzewo kategorii dla przykładu 1.



Rysunek 4.2: Drzewo kategorii dla przykładu 2.

W powyższych przykładowych drzewach $w(X, Y) = 1$, $w(Y, Z) = 2$, $D = 3$, stąd $d(X, Y) = \frac{1}{3}$, $d(X, Z) = \frac{2}{3}$. Miara wskazuje, że artykuły X i Y są do siebie mniej podobne, niż artykuły Y i Z .

4.1.2 Słowa kluczowe

NIE WIEM, CZY MIARA UŻYWAJĄCA KEYWORDÓW JEST POTRZEBNA

4.2 Miara 2: Ocena przez użytkowników offline

Kolejną wypracowaną miarą jest subiektywna ocena ekspercka. W celu obiektywizacji oceny ewaluacja powinna być dokonana przez grupę osób operujących na tych samych danych. Wadą tej metody jest jej powolność i potrzeba zaangażowania dodatkowych osób dokonujących ewaluacji. Niemożliwym wydaje się przeprowadzenie badania dla wszystkich artykułów, stąd konieczny jest wybór losowej próby artykułów, które parami poddane zostaną ocenie pod kątem podobieństwa. Skala ocen to 1-10: 1, gdy artykuły nie są do siebie podobne, 10, gdy podobieństwo jest całkowite.

4.3 Miara 3: Historyczna aktywność użytkowników serwisu

Zbieranie a następnie przechowywanie informacji o aktywności użytkownika w ramach serwisu internetowego jest powszechną praktyką. Proces ten pozwala na analizę zachowania użytkowników co może doprowadzić do wniosków, jakie usprawnienia należy przedsięwziąć, aby spełnić cele biznesowe. Jednym z przykładów aktywności użytkownika zapisywanej przez serwis Allego są kliknięcia w linki znajdujące się na stronie internetowej. Informacja ta pozwala sporządzić jeszcze jedną miarę jakości dopasowania podobnych do siebie artykułów. Postać danych, jakie udało mi się uzyskać z serwisu to tabela o polach: adres strony, na

której nastąpiło kliknięcie, adres strony, na którą prowadzi link, data kliknięcia.

Zaletą metody jest, iż można ją zastosować automatycznie, lecz jest zależna od danych analitycznych pochodzących z serwisu, które są niedoskonałe.

Jak już zostało opisane powyżej strona z artykułem tekstowym zawiera odnośniki do innych artykułów poruszających tematykę podobną do danego. Skoro zapisywana jest informacja o przejściach pomiędzy podstronami serwisu, to można policzyć ile razy z artykułu X dokonano przejścia na rekomendowany do niego artykuł Y_1 , a ile razy na rekomendowany artykuł Y_2 . Jeżeli liczba przejść na artykuł Y_1 jest większa niż na Y_2 , można wnioskować, iż Y_1 wydaje się być bardziej adekwatną rekomendacją dla artykułu X .

Posługując się powyższym założeniem, można zaproponować miarę jakości rekomendacji generowanych przez testowane metody w odniesieniu do popularności rzeczywistych rekomendacji wyekstrahowanej z danych serwisu o aktywności użytkowników.

W tym celu dokonuję adaptacji miary nDCG (Normalized Discounted Cumulative Gain). Miara ta służy do oceny jakości uszeregowania przedmiotów, np. wyników zwracanych przez silniki wyszukiwania.

TUTAJ OPISUJĘ MIARĘ + podaję źródło PROBLEM - algorytm allegro zmieniał się w czasie

Założmy, że dany algorytm A zwraca pewien ciąg artykułów $c_A = a_1, a_2, \dots, a_6$ podobnych do danego artykułu x , w kolejności od najbardziej adekwatnego. Założmy również, część elementów ciągu c_B artykułów rekomandowanych w serwisie dla x (używaną dotychczas w serwisie metodą B) znajduje się również w ciągu c_A , tj. np. ISTNIEJĄ TAKIE a_i, a_j (i, j to indeksy w ciągu c_A), że należą do c_A i c_B . Założmy ponadto, że z danych o kliknięciach użytkowników w linki w ramach serwisu wiadomo, że przejście z x na a_i jest bardziej popularne niż przejście z x na a_j . Stąd jeżeli $i < j$ ($i > j$), to jakość działania metody A jest dobra (zła), bo metoda ta generuje podobne artykuły w kolejności zgodnej ze stopniem podobieństwa z artykułem bazowym, opartym o częstość przejść użytkowników między artykułami.

Za wagi metody nDCG przyjmuje kiczby przejść pomiędzy artykułami, a

samą metodę stosuję tylko do przecięcia zbioru artykułów podobnych do danego generowanych przez daną metodę ze zborem artykułów rekomendowanych do danego przez dotychczasową metodę działającą w serwisie.

Rozdział 5

Opis i wyniki badań

5.0.1

model piaseckiego ile unikalnych słów on nie zawiera ile wystąpień słów nie zawiera - wykresy

5.1 Dalsze badania

Dalsze badania.

Niniejsza praca nie wyczerpuje sposobów wyboru artykułów podobnych.

Nie wszystkie pola zawarte w strukturze zostały wykorzystane: autor

Przed zastosowaniem metod wyznaczania podobieństwa wykonałem przetwarzanie wstępne dokumentów, które można przeprowadzić również na inne sposoby. Jest to temat osobnych badań.

Zdaję sobie sprawę z niedoskonałości zastosowanych miar...

Tematem niniejszej pracy jest przypisanie danemu artykułowi artykułów najbardziej podobnych. Warto tutaj zaznaczyć różnicę pomiędzy tematyką pracy a komercyjnym zagadnieniem najlepszych rekomendacji. Artykuły, które można uznać za dobre rekomendacje, tj. takie, które przynoszą przedsiębiorstwu największy zysk, wcale nie muszą być podobne do danego. Powszechnym zjawiskiem jest wzbogacanie rekomendacji o przedmioty niepodobne do danego,

a pozwalające użytkownikowi na poznanie osobnej kategorii przedmiotów, która może go zainteresować a tym samym przyciągnąć do serwisu.

Dodatek A

Technologie i narzędzie

Przy wykonywaniu operacji na tekście korzystałem głównie z silnika wyszukiwania Elasticsearch oraz własnoręcznie pisanych skryptów w języku Python wykorzystujących liczne specjalistyczne biblioteki posiadające interfejs w tymże języku. wypisać później użyte biblioteki

Bibliografia

- [1] Francesco Ricci, Lior Rokach, Bracha Shapira, *Introduction to Recommender Systems Handbook*, Springer, 2011
- [2] Słownik Języka Polskiego PWN <http://sjp.pwn.pl/sjp/arttykul;2441396.html> (07.05.2017)
- [3] <https://magazyn.allegro.pl/3333-serwis-allegro-to-nasz-sposob-na-was> (07.05.2017)
- [4] <http://morfologik.blogspot.com/> (07.05.2017)
- [5] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, *Efficient Estimation of Word Representations in Vector Space*, International Conference on Machine Learning (ICML), 2013
- [6] <https://code.google.com/archive/p/word2vec/> (26.05.2017)
- [7] <http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/> (26.05.2017)
- [9] Ofir Pele, Michael Werman, *Fast and robust earth mover's distances*, ICCV, 2009
- [10] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas, *The Earth Mover's Distance as a Metric for Image Retrieval*, str. 1, Computer Science Department, Stanford University, 2000

- [11] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas, *The Earth Mover's Distance as a Metric for Image Retrieval*, str. 8, Computer Science Department, Stanford University, 2000
- [12] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, Kilian Q. Weinberger, *From Word Embeddings To Document Distances*, International Conference on Machine Learning (ICML), 2015
- [13] https://en.wikipedia.org/wiki/Softmax_function/ (11.06.2017)
- [14] <https://allegro.pl/artykul/jaka-farba-dla-alergika-55917/> (26.06.2017)

Warszawa, dnia

Oświadczenie

Oświadczam, że pracę magisterską pod tytułem: „Rekomendacje artykułów opisujących produkty w serwisach e-commerce”, której promotorem jest dr inż. Anna Wróblewska, wykonałem samodzielnie, co poświadczam własnoręcznym podpisem.

.....