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

W niniejszej pracy zajmuję się porównaniem metod wyszukiwania podobnych do siebie artykułów tekstowych. Celem jest znalezienie w oparciu o treść danego artykułu podobnych artykułów, które możnaby zarekomendować użytkownikowi przeglądającemu dany artykuł. Problem zaczerpnięty jest z serwisu aukcyjnego Allegro, który posiada dział artykułów opisujących produkty dostępne w serwisie. Dział ten posiada system rekomendacji dopasowujący do danego artykułu listę artykułów, które są do niego najbardziej podobne i mogą zainteresować użytkownika. W swojej pracy staram się przeanalizować i zaaplikować znane metody wyszukiwania podobnych dokumentów tekstowych oraz porównać rezultaty. Skupiam się szczególnie na metodach opartych o semantyczną analizę tekstu.

Abstract

xcfghdfghhdfghfdgh

dfgdsfgsdfgsdfgds

Spis treści

Rozdział 1

Wstęp

1.1

Systemy rekomendacji są powszechnym elementem wielu serwisów internetowych. Sprawdzają się w 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. Pozwalato użytkownikowi na bardziej efektywne korzystanie z serwisu. Może to prowadzić do większego zaangażowania ze strony użytkownika i przywiązania do serwisu. Systemy rekomendacji dają obopulną korzyść zarówno użytkownikowi jak i właścicielowi serwisu internetowego.

Tematem mojej pracy magisterskiej jest stworzenie mechanizmu dopasowującego podobne do danego artykuły tekstowe w oparciu o ich treść. Szczegółowy problem poruszany w pracy pochodzi z serwisu Allegro posiadającego dział artykułów opisujących produkty dostępne w serwisie. W celu zachęcenia użytkownika do dalszej lektury artykułów stosuje się mechanizm rekomendacji podobnych artykułów. Celem niniejszej pracy jest zbadanie i udoskonalenie obecnego w serwisie mechanizmu generowania rekomendacji.

W swojej pracy korzystam z metod przetwarzania języka naturalnego a w tym z metod semantycznej analizy tekstu.

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

1.2 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 wcześciejsze zachowanie użytkownika.

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).

Od serwisu Allegro otrzymałem zserializowaną kopię 20000 artykułów dostępnych na stronach serwisu. Pojedynczy artykuł składa się z głównej zawartości tekstowej oraz pewnych metadanych. W celu otrzymania wszelkich danych od firmy Allegro wynagane było, abym podpisał umowę, w której zobowiązuję się

do nieujawniania żadnych danych, które otrzymałem. Stąd opisy danych, na których pracuję, zawarte w tej pracy nie wnikają w ich szczwgłóy i nieodbiegają od informacji publicznie dostępnych przez stronę allegro.pl.

W niniejszej pracy wykonuję eksperymenty wykorzystując znane metody określania podobieństw pomiędzy dokumentami, które adaptuję do zbioru dokumentów, które otrzymałem od serwisu Allegro.

W obszarze, którym zajmuje się niniejsza praca, bezpośrednim celem rekomendacji jest, aby użytkownik odwiedzał kolejne podstrony serwisu, co wprost zwiększa szansę na dokonanie przez niego transakcji.

Obecnie wykorzystywana metoda generowania rekomendacji artykułów opiera się o zapytanie do usługi Elasticsearch. Elasticsearch jest popularnym silnikiem wyszukiwania tekstu opartym o indeks Lucene. Działa w architekturze rozproszonej a komunikacja z nim następuje poprzez protokół HTTP i format JSON.

Metoda ta ogranicza się jednak jedynie do wyszukiwania tekstowego pomijając zagadnienia semantyczne. Znaczy to, że jeżeli dwa teksty opisują ten sam temat, ale używają to tego różnych słów, np. synonimów, to systemowi opartemu jedynie o wyszukiwanie tekstowe nie uda się stwierdzić podobieństwa między tymi tekstami, mimo, iż takowe istnieje.

Stąd w mojej pracy postanowiłem wykonać eksperymenty z metodami używającymi semantycznej analizy tekstu, aby ocenić, czy dają one lepsze rezultaty od obecnie stosowanej metody.

W niniejszej pracy skupiam się głównie na podejściu word2vec z racji tego, iż powstał niedawno.

Dochodzenie nowych rekomendacji - nie jest tematem pracy

napisać, że język polski stanowi trudność ——————-

Słowa, które zostają po tokenizacji to np literówki

1.3 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 wstępnego etapy przetwarzania ich, aby nadawały się do zaaplikowania do nich wybranych metod.

Dalej, w rozdziale 4 opisuję proces zastosowania wybranych metod oraz porównanie efektów ich działania. Opisuję również użyte metody ewaluacji wyników.

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

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 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 zainterezować. Działalność tekiego systemu daj zysk zarówn 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 przeferencje. 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

W celu formalizacji dalszych opisach stosowanych metod stosuję następujące

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 $bow : C \to \mathbb{R}^n$ 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 $i \in 1, 2, ..., n$ wektor reprezentujący dokument $d_i \in C$

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

Każdy dokument reprezentowany jest przez wektor, składający się z wag słów występujących w tym dokumencie. TFIDF informuje o częstości wystąpienia termów uwzględniając jednocześnie odpowiednie wyważenie znaczenia lokalnego termu i jego znaczenia w kontekście pełnej kolekcji dokumentów.

W celu sprowadzenia korpusu do reprezentacji bag-of-words

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 przypdku 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 frequecy" 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 sieci neuronowej, 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 siec przewiduje N sąsiednich słów lub CBOW: na postawie 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 $\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}\ dla\ j = 1,\ ..., K$. 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 trzenowaniu sieci neuronowe. 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 sieći, które faktycznie są wynikowymi wektorami słow.

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 słow 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 warstwy ukrytej sieci neuronowej. Wagi warstwy ukrytej można interpretować jako macierz MxN, 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 używa 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 opisanej powyżej metody jest utrata potencjalnie użytecznych zależności wektorami wchodzącymi w skład dokumentu.

W kontrze to tego prezentujemy metodę liczenia szukanego dystansu uwzględniającą rozkład wektorów wewnątz 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 dokumantu.

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: $P = \{f(p_1, w_{p_1})...(p_m, w_{p_m})\}$, $Q = \{f(q_1, w_{q_1})...(q_n, w_{q_n})\}$ 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 $Work(P, Q, F) = \sum_{i=1}^{m} \sum_{n}^{j=1} d_{ij} f_{ij}$ przo odpowiednich ogramiczeniach[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.

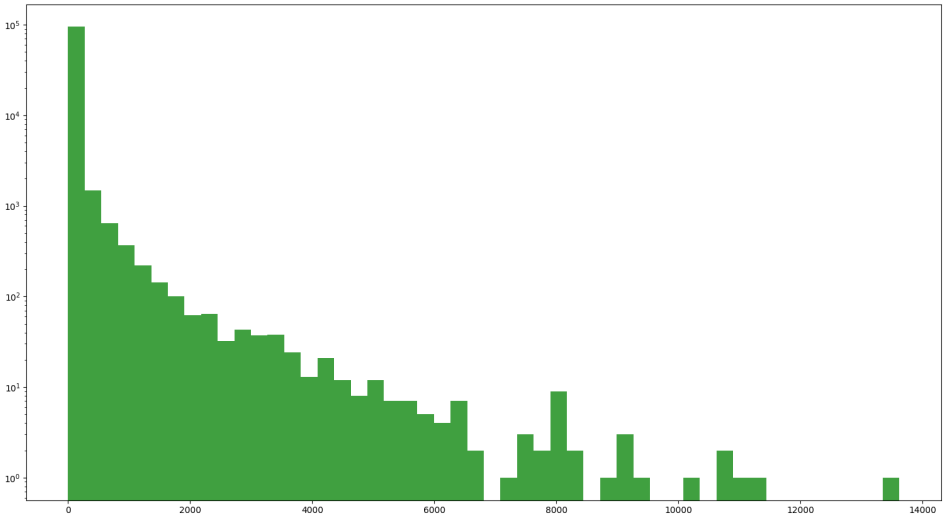
Rozdział 3

Dane

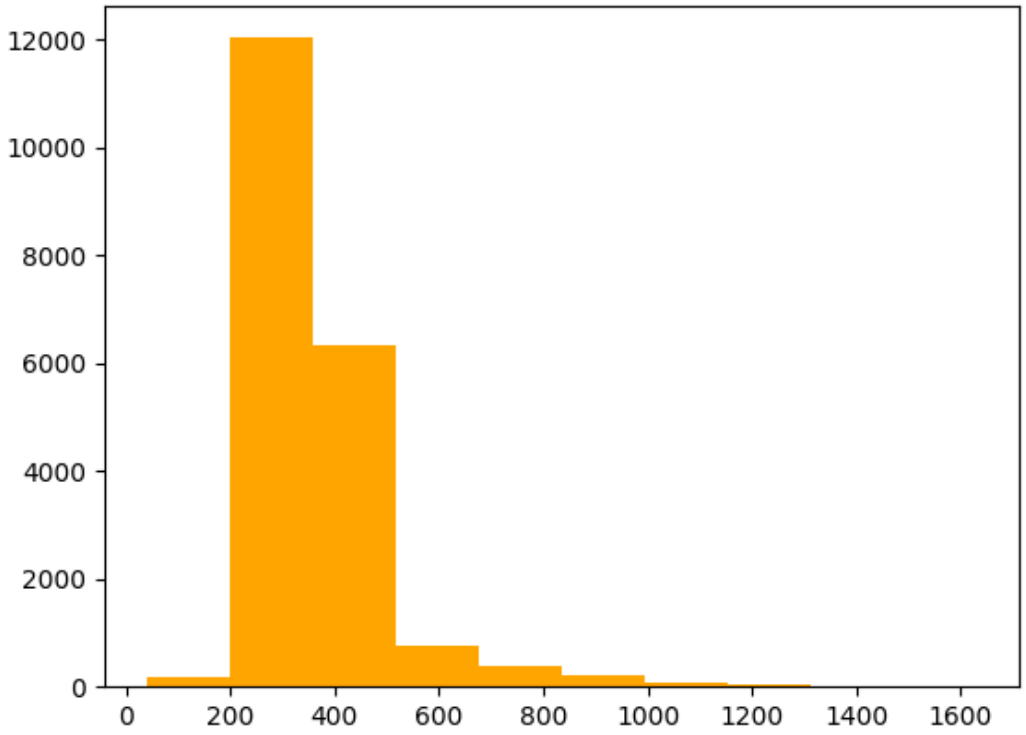
Dane, na których testowane były opisywane w niniejszej pracy metody otzrymałem dzięki życzliwości serwisu e-commerce Allegro. Jednak, by dane te otrzymać, zobowiązany zostałem po podpisania umowy o poufności. Stąd, w niniejszej pracy brak jakichkolwiek przykładów danych, a jedynie opisy metod użytych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

NApisać, że w korpusie znajduje się wiele specyficznych słów branżowych

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.



Rysunek 3.1: Histogram wystąpień słów w korpusie



Rysunek 3.2: Histogram długości artykułów.

3.1 Wstępne przetwarzanie danych

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

Surowe artykuły odtrymane od Allegro posiadają w swej treści wiele znaczników interpretowanych przez system, na podstawie których wzbogacana jest warstwa wizualna strony internetowej zawirającej 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 elemantem wstępnego pzetwarzania 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 etepach słownik zbudowan na korpusie zawiera 98174 słów.

filtracja ekstremalnych słów

3.2 Opis danych po wstępnym przetwarzaniu

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

W skład faktycznej treści artykułu wchodzą 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, ze kategorie są drzewiaste

Każdemu artykułowi 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.

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.

3.3 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 pewnyh założeniach, od których prawdziwości zlaeży jakość samej miary.

Działanie testowanych metod można sformalizwać w postaci pewnej funkcji $S : C \rightarrow ZAPYTAREJMERAjakzapisaskoczonycigelementwzC$. Funkcja S przyjmuje artykuł tekstowy (bądź jego identyfikator) i zwraca skończony ciąg artykłów 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. Oceny 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 stosuję średnią ważoną podobieństwa kolejnych elementów wyjścia z wejściem. Stosowane wagi: $\frac{1}{i}$ dla $i = 1, ..., N$, gdzie N to długość ciągu wyjściowego danej metody.

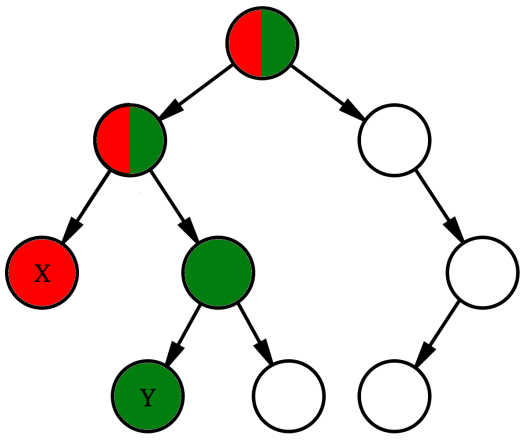
3.3.1 Miara 1: Dystans oparty na metadanych

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”.

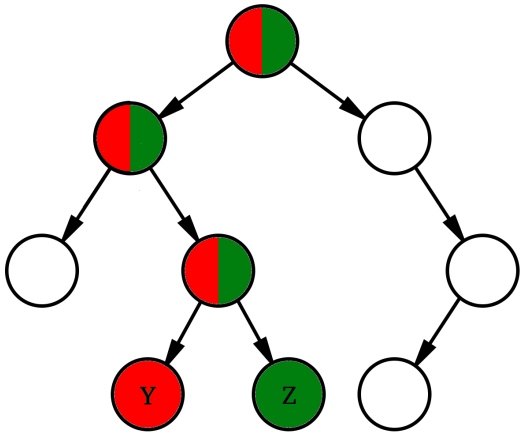
Kategorie

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 podobe 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 3.3: Drzewo kategorii dla przykładu 1.



Rysunek 3.4: Drzewo kategorii dla przykładu 2.

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

Słowa kluczowe

NIE WIEM, CZY MIARA UŻYWAJĄCA KEYWORDÓW JEST POTRZEBNA

3.3.2 Miara 2: Ocena użytkowników offline

Kolejną wypracowaną miarą jest subiektywna ocena ekspercka. W celu obiektywizacji oceny ewaluacja powinna być dokonana przez grupę osób operującycha 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.

3.3.3 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 odnośiki 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 Y1, a ile razy na rekomendowany artykuł Y2. Jeżeli liczba przejść na artykuł Y1 jest większa niż na Y2, można wnioskować, iż Y1 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

Załóżmy, że dany algorytm A zwraca pewien ciąg artykułów $c_A = a1,\ a2,\ ...,\ a6$ podobnych do danego artykułu x, w kolejności od najbardziej adekwatnego.

Załóżmy 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. Załóżmy 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 częstośc przejść użytkowników między artykułami. DO PRZEREDAGOWANIA!!!!

Za wagi metody nDCG przyjmuję 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.

3.4 Opis i wyniki badań

3.4.1

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

3.5 Dalsze badania

Dalsze badania.

Niniejsza praca nie wyczerpuje sposobów wyboru artykułów podobnych.

Nie wszystkie pola zawarte w strukturze zostały wykorzystane: autor

Przed astosowaniem 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 rekomandacje, tj. takie, które przynoszą przedsiębiorstwu największy zysk, wcale nie musż być podobne do danego. Powszechnym zjawiskiem jest wzbogacanie rekomandacji 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/artykul;2441396.html` (07.05.2017)

[3] `https://magazyn.allegro.pl/3333-serwis-allegro-to-nasz-sposob-na-wasze-s` (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)

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.

.....