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 dokonuję metod wyszukiwania podobnych dokumentów należących do zadanego zbioru artykułów. 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. W swojej pracy staram się przeanalizować i zaaplikować znane metody wyszukiwania podobnych dokumentów tekstowych. Skupiam się szczególnie na metodach opartych o semantyczną analizę tekstu.

Abstract

xcfghdfghhdfghfdgh

dfgdsfgsdfgsdfgds

Spis treści

Rozdział 1

Wstęp

Allegro jest największą działającą na rynku polskim platformą aukcyjną on-line. 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.

W celu zachęcenia użytkownika 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.

Tematem mojej pracy magisterskiej jest stworzenie mechanizmu dopasowującego podobne do danego artykuły tekstowe. Problem zaczerpnięty jest z serwisu Allegro, gdzie istnieje 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.

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.

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.

gtdfsgsdfgsdfg

Rozdział 2

Eksperymenty i testy

Rozdział 3

Przegląd znanych metod

W swojej pracy wykorzystuję i adaptuję do swoich potrzeb szereg metod i technik. Należą one do takich obszarów, jak: systemu rekomendacji, przetwarzanie języka naturalnego,

3.1 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).

3.1.1 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.

3.1.2 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.

3.2 Information retrieval

3.3 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: $V = w \mid \exists_{d \in C}\, w \in d$.

3.3.1 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 $b : C \to \mathbb{R}^n$ 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 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.

3.3.2 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.

3.3.3 Latent semantic indexing

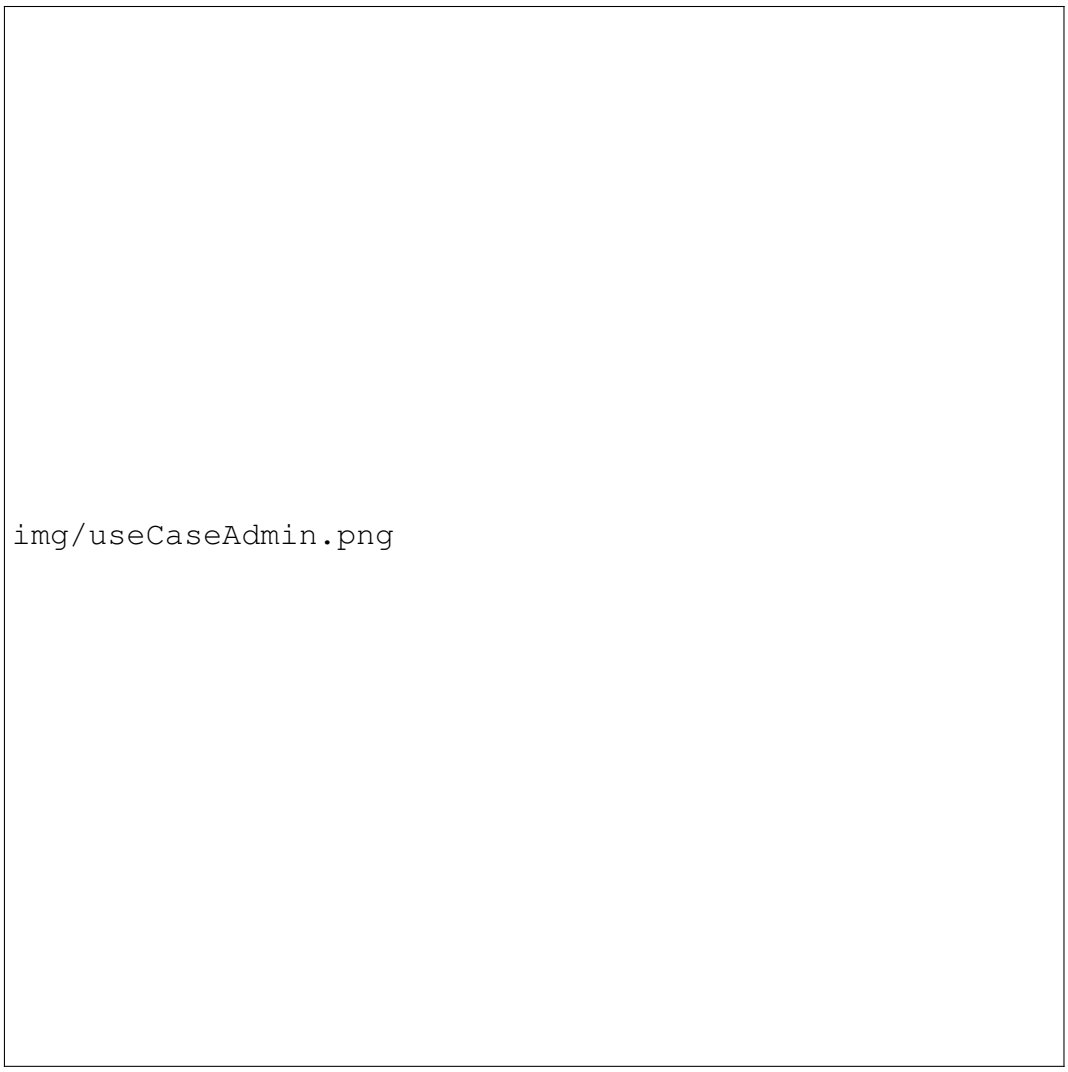
3.3.4 Latent Dirichlet allocation

3.3.5 Word2vec

3.3.6 Word mover's distance

3.4 Moduł administratora

Zawiera zestaw funkcjonalności związanych z zarządzaniem systemem oraz elementami, które pozostają widoczne na wszystkich podstronach.



Rysunek 3.1: Diagram przypadków użycia dotyczących zarządzania językami i menu

3.4.1 Edycja menu nawigacyjnego

Inicjator: administrator
Cel: umożliwienie szybkiego dostępu do najważniejszych elementów systemu
Główny scenariusz:

1. Administrator zgłasza chęć edycji menu nawigacyjnego.
2. Administrator wybiera spośród listy stron te, do których odnośniki mają znajdować się w głównym menu.

3. System zapisuje zmiany.

Rozdział 4

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.

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.

1. Warstwa dostępu do danych (Data Access Layer, DAL)

2. Warstwa logiki biznesowej (Business Logic Layer, BLL)

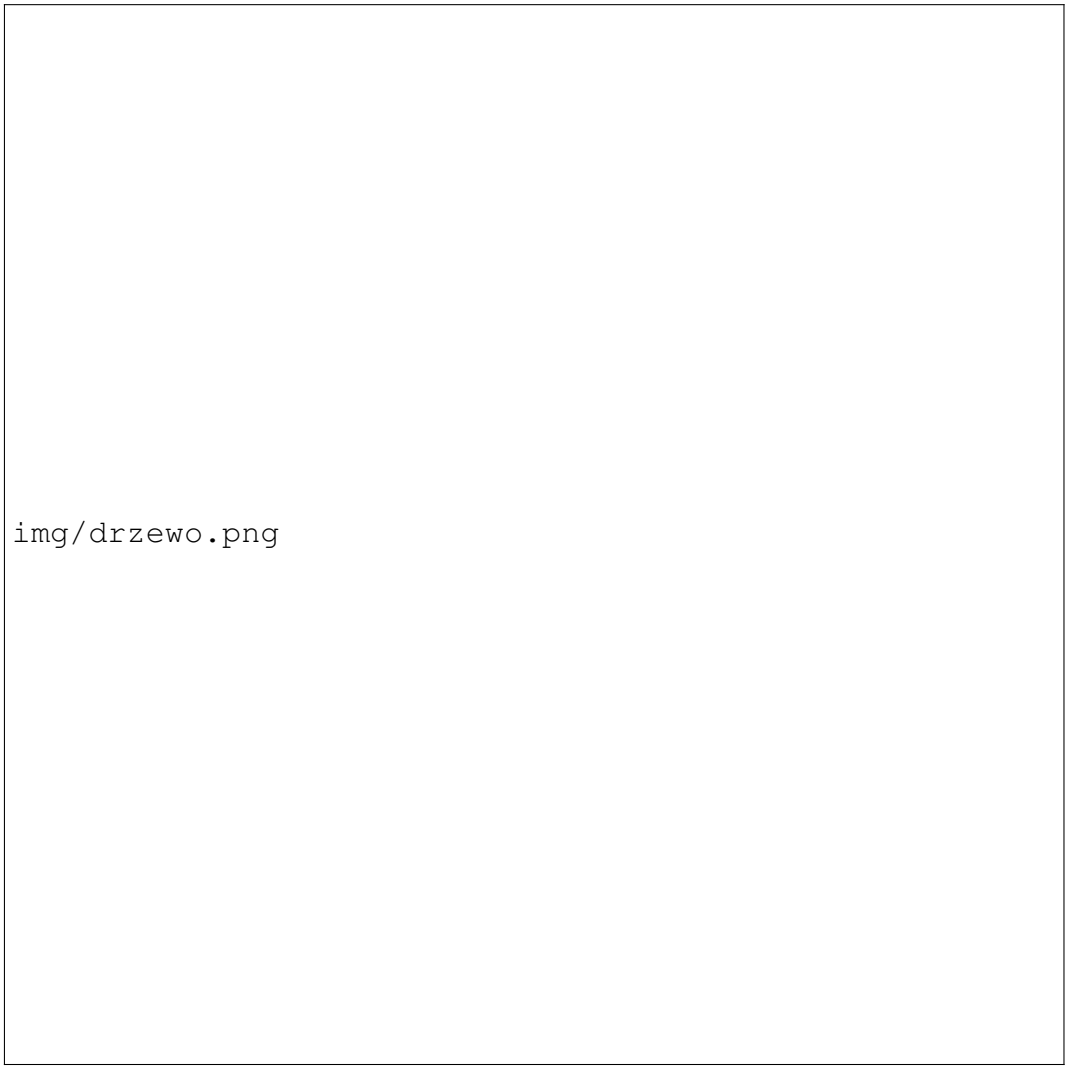
3. Warstwa interfejsu użytkownika (User Interface, UI)

    (a) Część serwerowa

    (b) Część kliencka

Rysunek 4.1: Diagram przedstawiający zarys architektury systemu

4.1 Projekt podstron



Rysunek 4.2: Diagram przedstawiający drzewo stron systemowych.

Rozdział 5

Słownik pojęć

W celu uniknięcia niejednoznaczności stosowanej w pracy terminologii definiujemy następujący słownik wykorzystywanych pojęć.

- Wymagania systemowe – zbiór wymagań jakie musi spełniać system operacyjny aby możliwa była poprawna praca systemu.
- Autoryzacja - kontrola dostępu, która potwierdza, czy dany użytkownik jest uprawniony do korzystania z żądanego zasobu.
- Konto – element systemu odpowiedzialny za przechowywanie podstawowych danych użytkownika systemu, jego uprawnień oraz roli pełnionej w systemie.

Bibliografia

[1] Francesco Ricci, Lior Rokach, Bracha Shapira, *Introduction to Recommender Systems Handbook*, Springer, 2011

[2] ASP.NET `http://www.asp.net` (12.01.2016)

[3] ASP.NET MVC `http://www.asp.net/mvc` (12.01.2016)

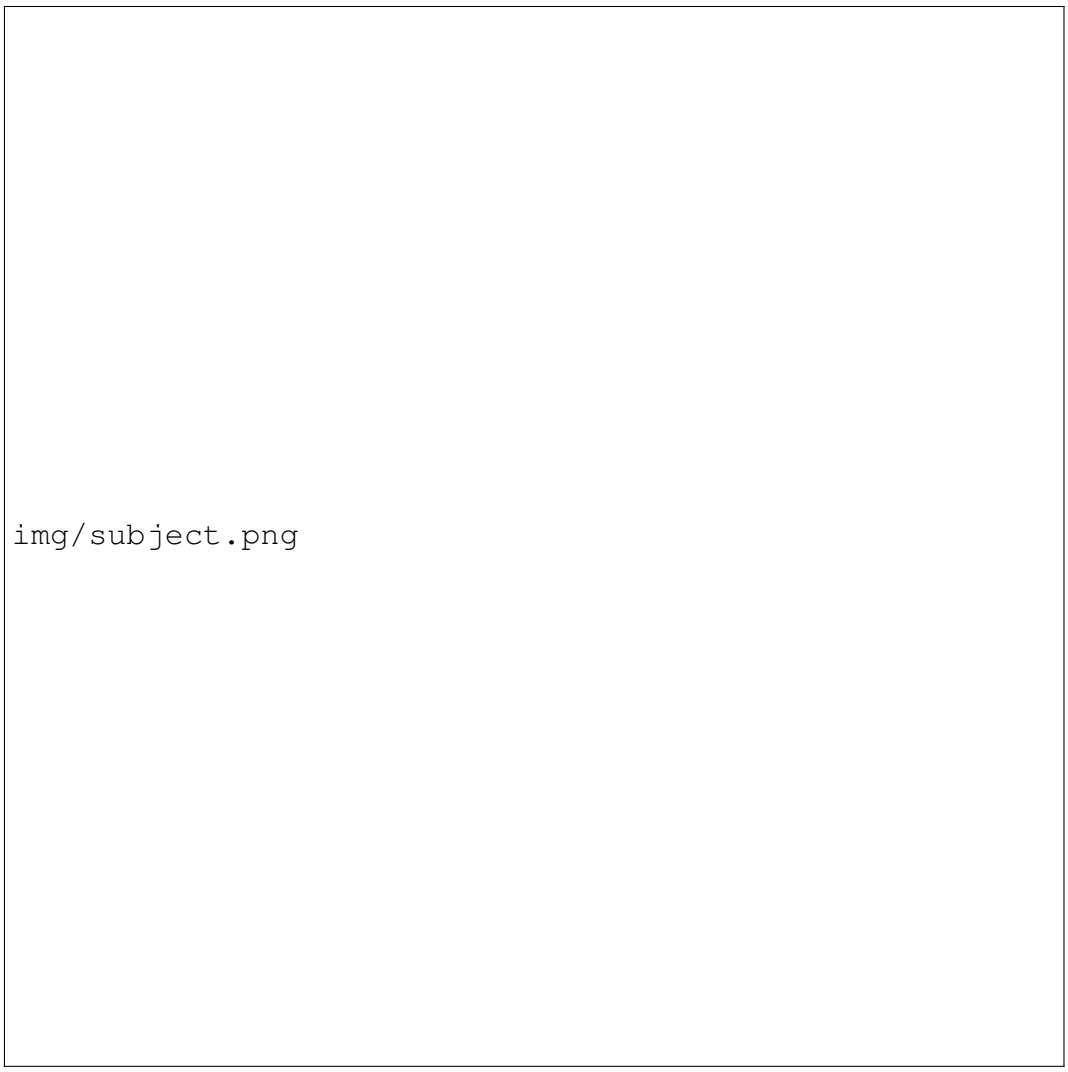
[4] IIS `https://www.iis.net/` (12.01.2016)

[5] SQL Server `https://www.microsoft.com/en/server-cloud/products/sql-server/default.aspx` (12.01.2016)

Dodatek A

Instrukcja użytkownika

Zalogowany nauczyciel ma możliwość zarządzania swoimi przedmiotami tzn. przedmiotami do których jest przypisany jako nauczyciel. Przejście do zarządzania przedmiotami następuje po wciśnięciu w menu głównym przycisku z loginem a następnie wybraniu „Moje przedmioty”. Po przekierowaniu istnieje możliwość dowolnej modyfikacji wyświetlanych treści a także akceptowania wniosków studentów o rejestrację a także podejrzenia listy studentów zapisanych na przedmiot.



Rysunek A.1: Ekran przykładowego przedmiotu

Dodatek B

Testy systemu

Oprócz testów jednostkowych dołączonych do kodu źródłowego aplikacji zostały przeprowadzone również testy funkcjonalne w środowisku produkcyjnym. Zakończyły się one sukcesem i udowodniły tym samym, iż system spełnia wymagania, które zostały postawione przed rozpoczęciem prac. W tym rozdziale za pomocą list kroków opisujemy przebieg kolejnych przypadków testowych.

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.

.....