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

1.1

Systemy rekomendacji są powszechnym elementem wielu serwisów internetowych. Sprawdzają się w takich polach jak polecanie produktów w sklepie czy rekomendacje ofert pracy. Pozwalają one użytkownikowi Dają użytkownikowi poczucie indywidualnego traktowania przez serwis internetowy dopasowujący niejako zawartość swoich stron to konkretnego użytkownika.Może to prowadzić do większego zaangażowania ze strony użytkownika i przywiązania do serwisu. Dają obopulną korzyść użytkownikowi oraz właścicielowi serwisu internetowego.

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 swojej pracy korzystam zarówno z metod wyszukiwania w teście

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 si głównie na podejściu word2vec z racji tego, iż powstał niedawno.

Dochodzenie nowych rekomendacji - nie jest tematem pracy ————————-

1.3 Struktura pracy

Rozdział 2 opisuje znane metody zaadaptowane przez mnie w eksperymentach.

Rozdział 2

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,

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 Information retrieval

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

2.3.1 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 $bow : C \to \mathbb{R}^n$ 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.

2.3.2 Term frequency - inverted document frequency

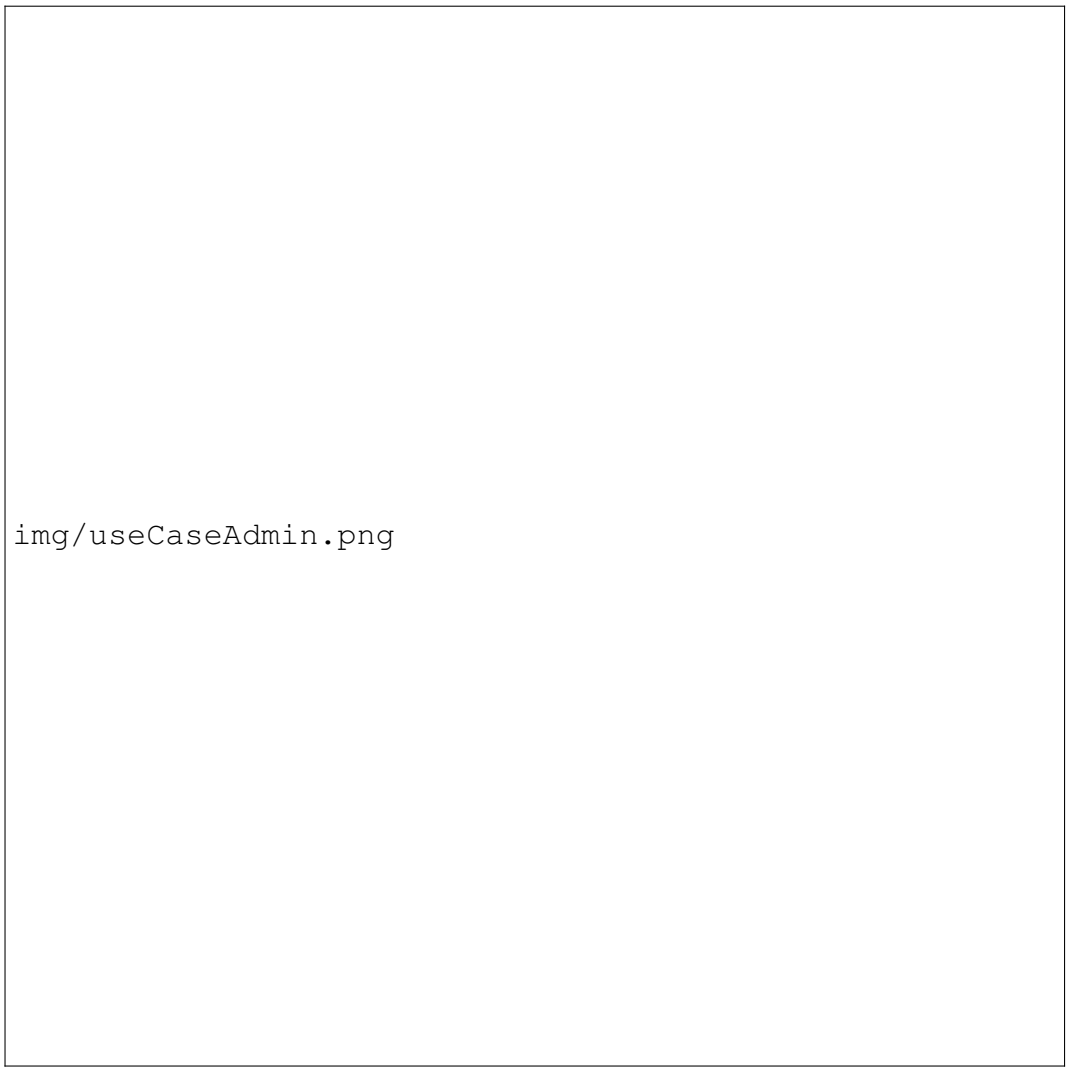
TF-IDF(ważenie częstością termów - odwrotna częstość w dokumentach) jest metodą reprezentacji tekstu jako zbioru słów przy jednoczesnym uwzględnieniu wagi słów, która zależy od częstości występowania słowa w korpusie. Oznaczenia formalne takie same tak w przypdku BOW. $v_{ij} = tfidf_{ij} = tf_{ij} *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

2.3.3 Latent semantic indexing

2.3.4 Latent Dirichlet allocation

2.3.5 Word2vec

2.3.6 Word mover's distance



Rysunek 2.1: Diagram przypadków użycia dotyczących zarządzania językami i menu

2.3.7 Edycja menu nawigacyjnego

1. Administrator zgłasza chęć edycji menu nawigacyjnego.

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

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 maluk 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.

3.1 Metody ewaluacji

Ewaluacja

W celu porównania stosowanych metod wyznaczania podobieństwa między artykułami konieczna jest formalizacja pewnej miary tego podobieństwa.

Ewaluacja rankingu, którym niewądfgdg .. 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.

Miara 1: jak daleko pod względem kategorii jesteśmy

Pierwszą zastosowaną miarą, pozwalającą ocenić jakość dopasowania

podobnych artykułów jest ich odległość ww wcześniej wspomnianym drzewie kategorii: im mniejszy dystans pomiędzy liśćmi drzewa, tym większe podobieństwo pomiędzy artykułami. 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 obiektywnie uzasadniona. Kolejną zaletą tej miary (w odróżnieniu od następnej) 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.

Miara 2: subiektywna - trzeba wymyślić jakąś punktację

Kolejną wprowadzoną miarą jest subiektywna ocena ekspercka. W celu obiektywizacji oceny ewaluacja powinna być dokonana przez więcej niż jedną osobę. 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 poddane zostaną ocenie.

Miara 3: kliki

3.2 Wyniki badań

3.3 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 muszą 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.

Rozdział 4

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] Słownik Języka Polskiego PWN `http://sjp.pwn.pl/sjp/artykul;2441396.html` (07.05.2017)

[3] Allegro `https://magazyn.allegro.pl/3333-serwis-allegro-to-nasz-sposob-na-w` (07.05.2017)

[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

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.

.....