Uniwersytet Kardynała Stefana Wyszyńskiego
w Warszawie
Wydział Matematyczno-Przyrodniczy
Szkoła Nauk Ścisłych

Jacek Giedronowicz
Nr albumu: 95175

Rekomendacja stron www za pomocą silnika Apache Lucene

Praca licencjacka na kierunku *Informatyka*
w zakresie *Machine Learning*

Praca wykonana pod kierunkiem
dr. Roberta Kłopotka

Lipiec 2020

Słowa kluczowe

-

Dziedzina Socrates-Erasmus

11.3 Informatyka

Klasyfikacja tematyczna

Machine Learning

English title

Website recommendation using the Apache Lucene engine

Spis treści

Rozdział 1

Wprowadzenie

- ok 1-2 str
- ogólne zagadnienie jak jest stosowane w praktyce (do przewidywania)
- opis, że klasyfikacja jest trudna, używa się wielu metod, że jednym z podejść jest komitet klasyfikatorów, aby poprawić skuteczność
- cel pracy, hipotez badawczych
- krótki opis co jest w danym rozdziale

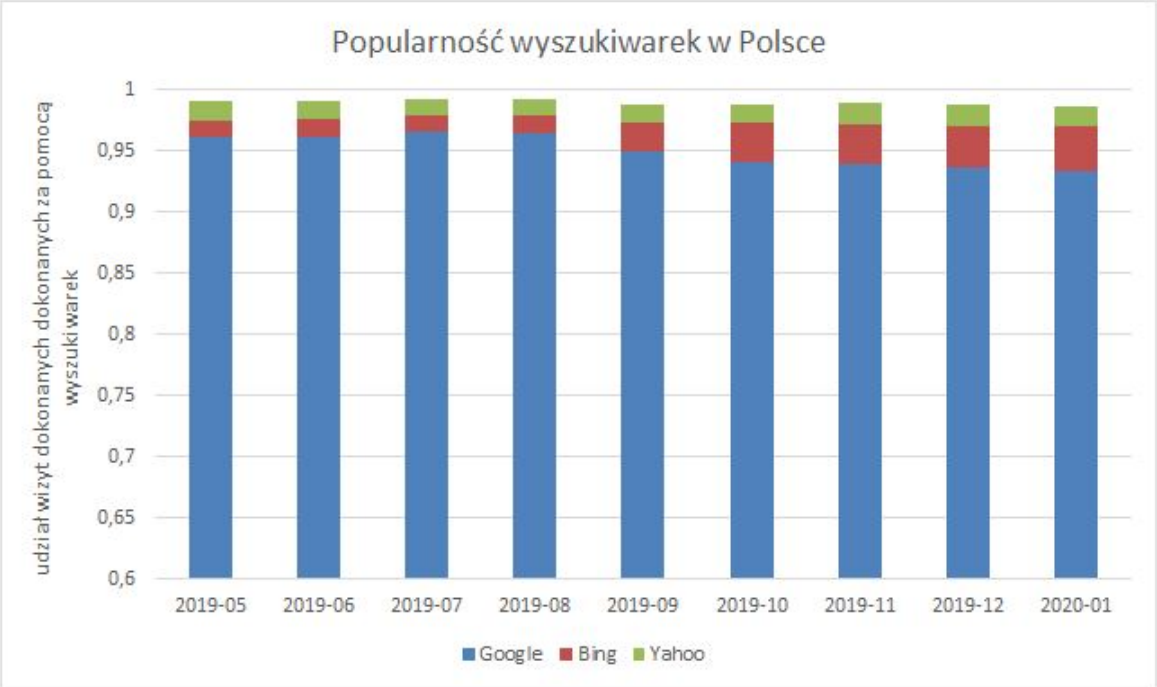
Rozdział 2

Stan rynku

- przede wszystkim zdefiniowanie problemu biznesowego i jego uzasadnienie.
- Jakie są wyszukiwarki?
- Jakie mają opcje?
- strony spamerskie (zaburzają ranking)
- hint: zobacz prace lic o serwisie otomoto

Na rynku istnieje wiele wyszukiwarek internetowych, lecz mało kto potrafi wymienić więcej niż pięć. Internet zdominowała firma Google. Nawet takie znane wyszukiwarki jak Bing od Microsoft'u czy Yahoo mają zaledwie kilka procent udziału, który przedstawia się następująco:

- Google - 93,37%
- Bing - 3,61%
- Yahoo - 1,75%
- DuckDuckGo - 0,29%
- Interia Katalog - 0,14%



Rysunek 2.1: Popularność wyszukiwarek w Polsce
Źródło: gs.statcounter.com – styczeń 2020

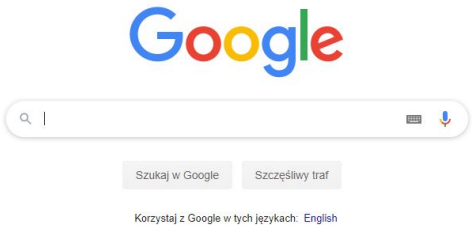
W Polsce, z powodu trudności w przetwarzaniu naszego języka przez zagraniczne systemy, były próby stworzenia wyszukiwarek dedykowane dla polaków. Systemy z algorytmami rozpoznawania odmian słów w języku polskim. Największe powodzenie miała witryna szukacz.pl, która funkcjonowała przez 10 lat od 2001 do 2011 roku. Jak możemy przeczytać na stronie http://szukacz.pl/wiecej_o_szukaczu.html [3]:

> "W szczycie swojego rozwoju, pod koniec 2007 roku, odpowiadał ze 115 milionów dokumentów w języku polskim pochodzących z miliona witryn (kolekcja „Polska"). Do połowy 2007 roku odpowiadał także z 45 milionów wyselekcjonowanych dokumentów w języku angielskim pochodzących z 2 milionów witryn (kolekcja „Świat"; tylko 4 procent pytań było skierowane do tej kolekcji)."

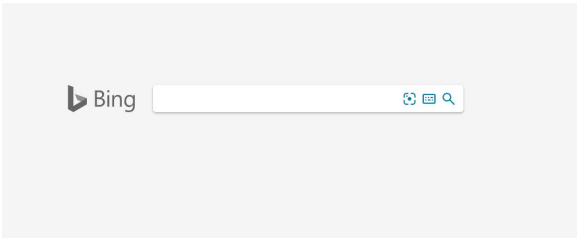
Mimo sporej bazy dokumentów oraz przyzwoitym budżecie, również i ta witryna musiała się poddać międzynarodowemu gigantowi jakim jest Google. Stąd też wybrałem witrynę google.pl jako wzór, do którego będę się odnosił przy implementacji własnego systemu. W dalszej części opiszę najważniejsze funkcjonalności witryny Google i porównam z Bing.

2.1. Wyszukiwarka

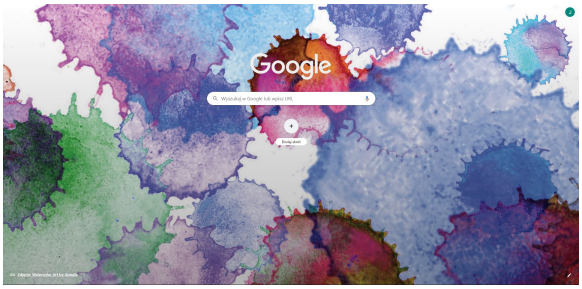
Na rysunkach 2.2 i 2.3 przedstawiono główne strony wyszukiwarek Google i bing. Jak widać obie strony są minimalistyczne i intuicyjne. Witryny oferują również możliwość jej spersonalizowania przez bogatą bazę skórek, co zostało przedstawione na rysunkach 2.4 i 2.5



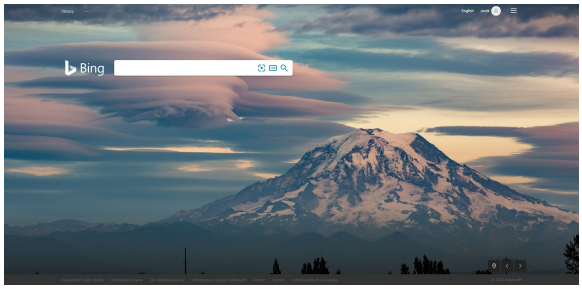
Rysunek 2.2: Główna strona wyszukiwarki google



Rysunek 2.3: Główna strona wyszukiwarki bing



Rysunek 2.4: Główna strona wyszukiwarki Google ze skórką



Rysunek 2.5: Główna strona wyszukiwarki Bing ze skórką

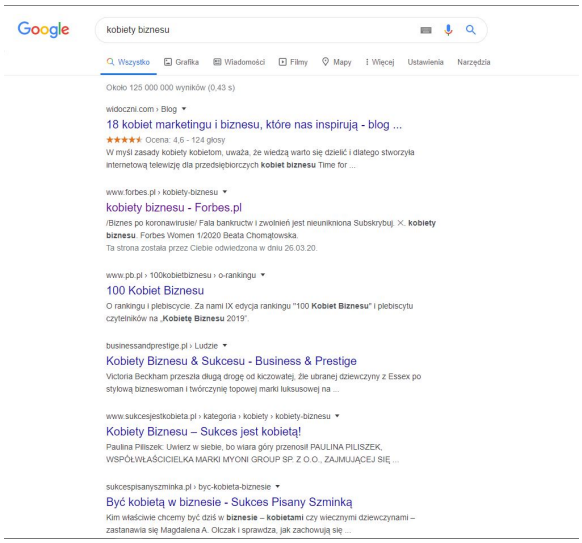
2.2. Wyszukiwanie

Po wpisaniu frazy, którą chcemy wyszukać obie wyszukiwarki wypiszą nam listę od kilkudziesięciu tysięcy do nawet kilkuset milionów rekomendowanych stron dla tego hasła.

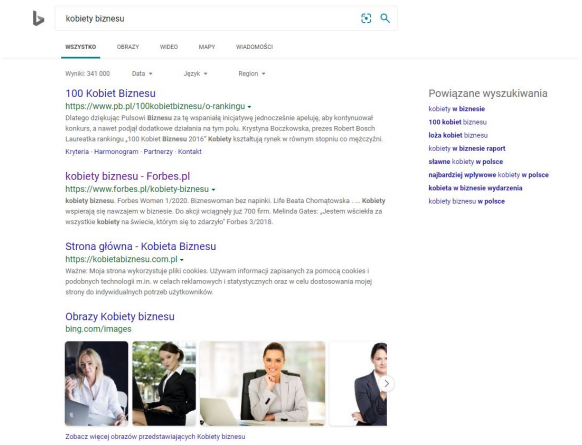
Dla przykładu dla frazy „kobiety biznesu" Google znalazło 125 000 000 wyników kiedy Bing zaledwie 341 000. Dominacja wyszukiwarki Google wśród użytkowników oraz zdecydowanie większa pula zaindeksowanych stron może być powodem dlaczego współcześni deweloperzy przygotowują swoje strony głównie pod pozycjonowanie wyszukiwarki Google.

Każda z wyświetlanych pozycji przekazuje nam podstawowe informacje takie jak:

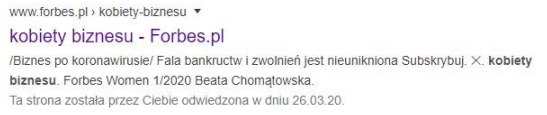
- tytuł strony (zaznaczony kolorem zielonym)
- link do strony (zaznaczony kolorem fioletowym)
- skrót zawartości strony (zaznaczony kolorem brązowym)



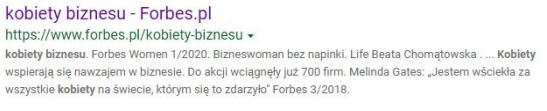
Rysunek 2.6: Lista rekomendowanych przez Google stron dla hasła: kobiety biznesu



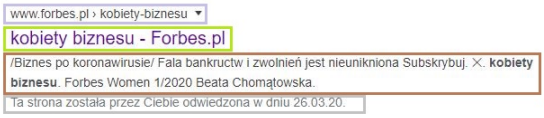
Rysunek 2.7: Lista rekomendowanych przez Bing stron dla hasła: kobiety biznesu



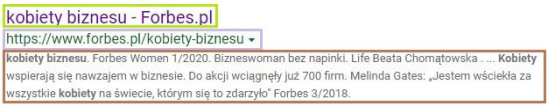
Rysunek 2.8: Wyświetlana pozycja w liście wyników google



Rysunek 2.9: Wyświetlana pozycja w liście wyników bing



Rysunek 2.10: Wyświetlana pozycja w liście wyników Google z zaznaczonymi obszarami

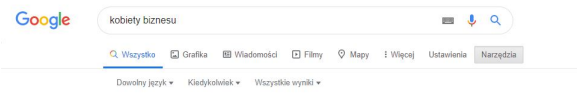


Rysunek 2.11: Wyświetlana pozycja w liście wyników Bing z zaznaczonymi obszarami

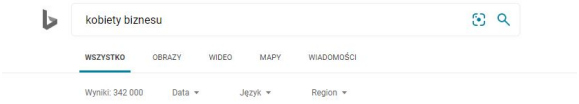
Oprócz podstawowego wyszukiwania linków do stron, mamy możliwość wybrania kategorii przedstawione na rysunkach 2.12 i 2.13 takich jak:

- grafika
- wiadomości
- filmy

Ponadto wyszukiwarki dają nam możliwość intuicyjnego filtrowania wyników. W tym punkcie wyszukiwarki nieznacznie się różnią, gdyż Bing oferuje takie parametry jak: data, język, region. Natomiast Google: data, język i możliwość pokazania dokładnych wyników zawężając o wyniki fraz bliskoznacznych.



Rysunek 2.12: Kategorie i filtry wyszukiwania Google



Rysunek 2.13: Kategorie i filtry wyszukiwania Bing

2.3. Podstawowe operatory

Aby podnieść jakość i zawęzić pole poszukiwań pożądanych informacji przez użytkownika, Google zaimplementowało tak zwane operatory. Operatory definiują zakres poszukiwań poprzez słowa i znaki kluczowe:

- *" "* - wpisując wyrażenie w cudzysłów wymusza ścisłe dopasowanie. Przykład: fraza *"Steave Jobs"* wpisana z cudzysłowem wymusza aby w wyniku były dokładnie te dwa słowa obok siebie.
- *OR* - operator logiczny pozwalający z zapytanie *Steave OR Jobs* wyszukać wyniki zawierające frazę *Steave* lub *Jobs* lub oba. Operator można zastąpić znakiem |.
- *AND* - domyślny operator logiczny, wyświetlający tylko te wyniki, które zawierają frazę *Steave* i *Jobs*. Operator jest domyślny więc przydatny jest tylko w połączeniu z innymi operatorami.
- *−* - operator wykluczający. W przykładzie *jaguar speed -car* wynikiem będą strony powiązane z frazą *jaguar speed* ale niezawierające *car*. Zatem chodziło nam o znalezienie prędkości jaguara będącego zwierzęciem a nie samochodem.
- *..* - operator zakresu liczb. Przykład: *podatki 2010..1015* da nam wynik wyszukiwania związanymi z podatkami miedzy 2010 a 2015 rokiem.
- *( )* - operator grupowania fraz z innymi operatorami. Przykład: *(iPad OR iPhone)apple*. Operator *AND* jest domyślny więc nie jest obowiązkowy.
- *define:* - operator wyszukuje definicje szukanej frazy.
- *filetype:* lub *ext:* - operator określający rodzaj pliku będący wynikiem wyszukiwania takie, jak pdf, docx, txt, ppt. Przykład: *Steve Jobs ext:pdf* - wyświetli tylko pliki pdf związane ze Steve'em Jobs'em.
- *site:* - ogranicza wyniki wyszukiwania tylko do jednej strony lub domeny. Przykład: *wniosek site:.gov.pl* wyszuka informacje o wnioskach tylko na stronach rządowych kończących się na *.gov.pl*.
- *inurl:* - operator przeszukuje podaną frazę tylko w adresie url strony.
- *intitle:* - operator przeszukuje podaną frazę tylko w tytule strony.
- *intext:* - operator przeszukuje podaną frazę tylko w treści strony. Pominie strony zawierające frazę w tytule czy adresie url.

Google zawiera znacznie więcej operatorów. Niektóre z nich wymienione są na stronie sites.google.com [2]

Rozdział 3

Rekomendacja stron www

3.1. Algorytm

Cały algorytm możemy podzielić na dwa etapy: indeksowanie i wyszukiwanie. Użytkownik oczekuje od aplikacji szybkiego i trafnego wyszukiwania jednak, aby tak się stało trzeba najpierw zebrać wszystkie dane i nadać im specjalną strukturę dzięki której będziemy mogli szybko znaleźć interesującą nas informacje. To właśnie nazywamy procesem indeksowania.

3.1.1. Proces indeksowania

Document

Analyzer

Struktura indeksu

3.1.2. Proces wyszukiwania

3.2. Dodatkowe funkcje składni zapytania

Rozdział 4

Przedstawienie problemu i sposób jego rozwiązania

- Przede wszystkim przedstawienie tego problemu, który będzie opisany w pracy
- problem przedstawiony w pracy może być jednym z podproblemów ogólnego problemu biznesowego
- opis metod lub algorytmów np w postaci pseudokodu lub schematów
- uzasadnienie użytych metod
- testy kilku parametrów, przygotowanie danych itd.

Rozdział 5

Funkcjonalność zaimplementowanego systemu

- opis jakie funkcjonalności będzie miał system
- w jaki sposób dana funkcjonalność przyczyni się do rozwiązania problemu

---

Rozdział 6

Implementacja oraz użyte technologie

- dokładnie jak zostało to wszystko implementowane
- opis architektury
- jaka platforma, diagram klas, jaka funkcjonalność jest w jakiej klasie/module/package
- opis głównych metod (nie opisujemy wszystkich metod)
- parametry jakie ma aplikacja, opis instalacji

6.1. Java

6.1.1. Apache Lucene

6.2. Spring Boot

6.3. Thymeleaf

6.4. HTML

6.5. Struktura i działanie klas programu

6.5.1. Klasy silnika wyszukiwania

6.5.2. Klasy kontrolera

Rozdział 7

Przypadki użycia

- co dokładnie trzeba zrobić, aby zrealizować daną funkcjonalność
- zrzuty ekranu

Rozdział 8

Podsumowanie

- ok 1-2 str
- cel pracy udało się osiągnąć
- w jaki sposób udało się osiągnąć cel z rozdziału 1 (w których rozdziałach jest to opisane)
- napotkane problemy
- w jaki sposób można aplikację/system udoskonalić

---[1]

Bibliografia

[1] Google power searching. *British Columbia Institute of Technology*, 2007.

[2] Google. Google search education evangelism. `https://sites.google.com/site/gwebsearcheducation/advanced-operators`, 2011.

[3] 24 Godziny Sp. z o.o. szukacz.pl. `http://szukacz.pl/wiecej_o_szukaczu.html`, 2000-2018.