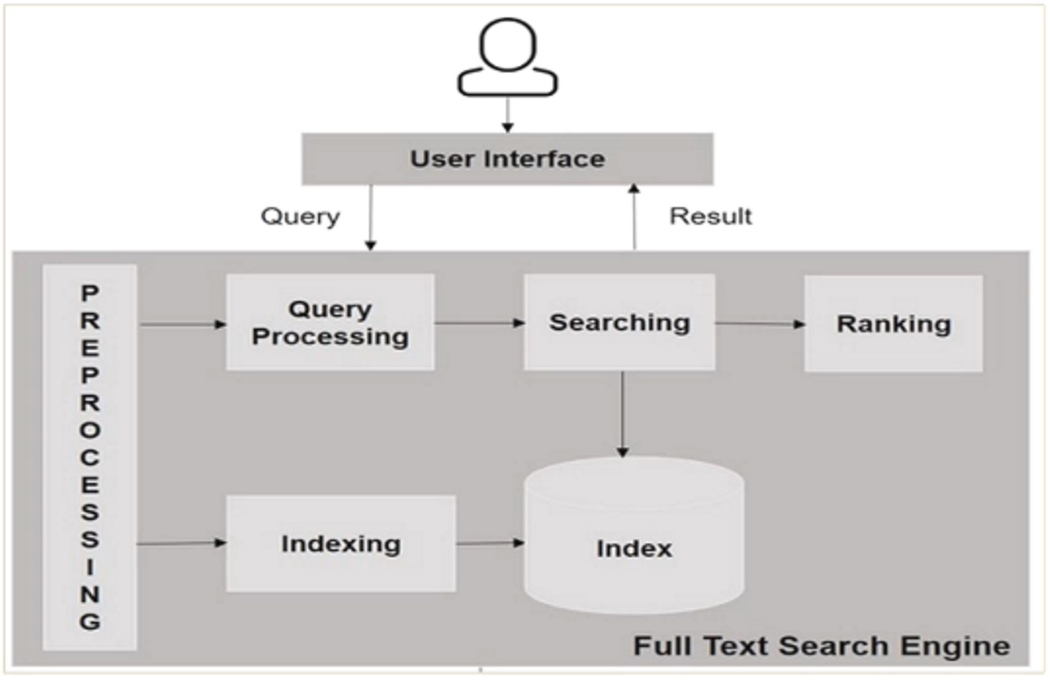
6. Full-text pretraga informacija

Sunday, 20 February 2022 16:09

Information retrieval

Najpoznatiji sistem za pretragu informacija pretstavljaju internet pretrazivaci (google)
Radi na osnovu poredjenja upita korisnika sa netekstuiranim dokumentima I vracanje upita sortiranog po relevantnosti

Zbog velicine podataka imamo razne tehnike pretrazivanja
- Tehnike:
  - Bazirana na pretrazi metapodataka dokumenta
    - Pretrazuju se samo meta podaci ne sadrzaj dokumenta
  - Bazirana na pretrazi full-text indeksima (indeksima kreiranim na osnovu sadrzaja)
    - Pretrazuje se kompletan sadrzaj dokumenta
    - Bitni aspekti pretrage
      - Relevantnost -
      - Analiza - pretvara blok teksta u tokene (token je celina, npr rec)
  - Mesavina ove dve



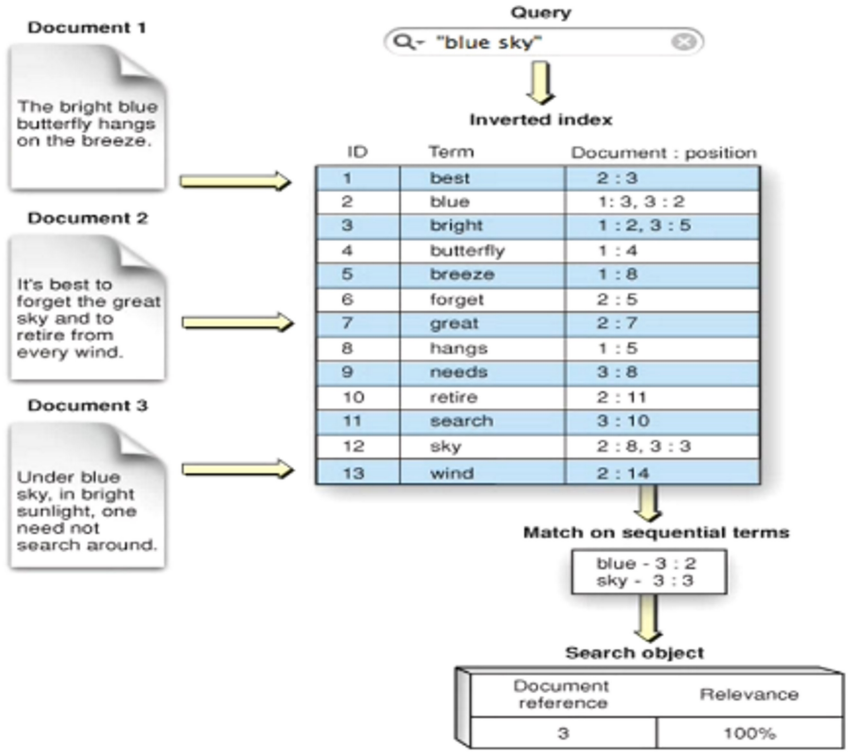
- Preprocesiranje
  - Kroz ovo prolaze I dokumenti I korisnicka pretraga
  - Ovo je samo analiza podataka kako bi se indeksirali
  - A radi se I u korisnickoj pretrazi kako bi se sveli na iste mere I olaksali pretragu
    - Kao svodjenje svega na mala slova itd.
- Posle se rankira I vraca korisniku rezultat

- Dodatne transformacije mogu da se dese kao ako pogresno spelujemo rec ispravice se pa ce se potraziti

Invertovani indeks

- Struktura koja se koristi za pretragu full-text
  - Vokabular - lista svih reci koje se pojavljuju u dokumentu

Pojavljivanja - lista koja sadrzi info o svakoj reci vokabulara
Gde u dokumentu se pojavljuje rec
Koliko puta se pojavljuje rec
Dokument u kome se pojavljuje rec



Rec blue se pojavljuje u dokumentu 1 na poziciji 3

Pretrazivanje
Rezultat upita je ono gde se pojavljuju obe reci u tekstu
100% je ovde al moze da bude I manje kao za dokument 1 gde se blue javlja jednom tj 50% relevantnosti

Lucene indeks je najpopularnija biblioteka za invertovani indeks
Java jezik

Elasticsearch
Platforma za skladistenje I pretragu dokumenta bazirana na Lucene indeks-u
Pretraga sadrzaja dokumenta
Implementirano Java, skladisti info kao JSON

Pojmovi:
Klaster - kolekcija jednog ili vise servera zaduzeni da zajedno cuvaju podatke I indeksuju ih I omoguce pretragu
Cvor - server na kome se cuvaju podaci
Indeks - kolekcija dokumenta koje imaju neke zajednicke karakteristike
Tip - Logicko particionisanje indeksa

Dokument - osnovna informativna jedinica (json)
Indeksiranje dokumneta - pribavljanje, obrada, analiza dokumenta I azuriranje indeksa
Shard - velike kolicine podataka se dele na male delove, nalaze se na raznim nodovima u klasteru
Svaki shard je nezavisan, ako otkaze jedan sard te informacije su izgubljene al ostali cvorovi koji rade su nezavisni pa nastave da rade
Replike - svaki shard moze da ima jednu ili vise replika zbog otkaza

TERMINOLOGY

MySQL	Elastic Search
Database	Index
Table	Type
Row	Document
Column	Field
Schema	Mapping
Index	Everything is indexed
SQL	Query DSL
SELECT * FROM table …	GET http://…
UPDATE table SET …	PUT http://…

Analizator ako ne navedemo dobijamo standardni

- Menja sva slova u mala
- Izbacuje nepotrebne reci kao clanovi, veznici itd.

Ima mnogo analizatora