Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Luka Dragar, Matej Vatovec, Aleksander Kovač

Inverted Index

Domača naloga pri predmetu Iskanje in ekstrakcija podatkov s spleta
MENTOR: prof. dr. Marko Bajec, doc. dr. Slavko Žitnik

Pri domači nalogi smo obdelali dano besedilo v večih dokumentih in dobljene podatke poindexirali. Najprej smo besedilo izluščili ter preobdelali, tako da smo ga tokenizirali, kasneje še normalizirali (izločili ločila, številke in izvedli lematizacijo ter besede shranili shranili v indeksno strukturo, s čimer smo poskrbeli za učinkovito strukturo iskanja in poizvedovanja po besedilu znotraj HMTL datotek.

• Indeksiranje podatkov

Program se sprehodi čez vsako HTML datoteko posebaj in naredi sledeče:

- izlušči samo besedilo (brez HTML značk...)
- tokenizira izluščeno besedilo
- kopira dobljeno tabelo tokenov in jo normalizira:
    - celotno besedilo spremeni v male črke
    - sprehodi se čez vsak token:
        - če je token ločilo ga zamenja z PUNCT
        - če je token številka (ali vsebuje številko) ga zamenja z NUM
        - če je token stopword ga zamenja z STOPWORD

tako dobimo dve tabeli. Primer:

```
['Galerija', 'MGML', ',', 'ki', 'se', 'nahaja', 'v' 'Ljubljani', ',', 'se', 'zapira', '.']
['galerija', 'mgml', 'PUNCT', 'STOPWORD', 'STOPWORD', 'nahajati', 'STOPWORD', 'ljubljana',
'PUNCT', 'STOPWORD', 'zapirati', 'PUNCT']
```

Prva tabela je tokenizirano besedilo, druga tabela pa normalizirano besedilo. Obe tabeli sta iste dožine zato da imajo vse besede iste indekse, kar olajša iskanje originalnega besedila pri poizvedbi. Pri indeksiranju se uporablja le druga tabela, torej normalizirano besedilo

Progam se sprehodi čez vse tokene v normalizirani tabeli (druga tabela) in si za vsako besedo shrani število pojavitev in indekse posamezne pojavitve. Pri tem ignorira tokene PUNCT, STOPWORD in NUM.

Dobljene besede in njene pojavitve skupaj z imenom trenutne datoteke spravi v podatkovno bazo.

• Iskanje podatkov brez obrnjenega indeksa

Programu preko ukazne vrstice podamo parametre. Ti se tokenizirajo in normalizirajo po istem postopku, kot je opisan v prvem poglavju. Program se nato sprehodi čez vse HTML datoteke in iz vsake izlušči samo besedilo. Nato se dobljeno besedilo tokenizira in normalizira, kot je opisamo v poglavju *Indeksiranje podatkov.* Sprehodimo se čez vsak token v normalizirani tabeli in pogledamo če se ta nahaja v podanih parametrih. Če se povečamo število njegovih pojavite in shranimo njegov indeks (položaj tokena v tabeli). Ko imamo preštete vse pojavitve in shranjene indekse, lahko s pomočjo tokenizirane tabele in indeksov pridobimo originalni tekst ter ga izpišemo.

• Iskanje podatkov z obrnjenim indeksom

Vzpostavi se povezava z podatkovno bazo, kjer se nahaja indeks. Programu prav tako, kot v prejšnjem primeru preko ukazne vrstice podamo parametre. Ti se tokenizirajo in normalizirajo po istem postopku, kot je opisan v prvem poglavju. Nato se naredi poizvedba za vsako besedo v podanih parametrih:

```
SELECT documentName, sum(frequency), group_concat(indexes) FROM (" + subquery
+ ") GROUP BY documentName ORDER BY sum(frequency) DESC
```

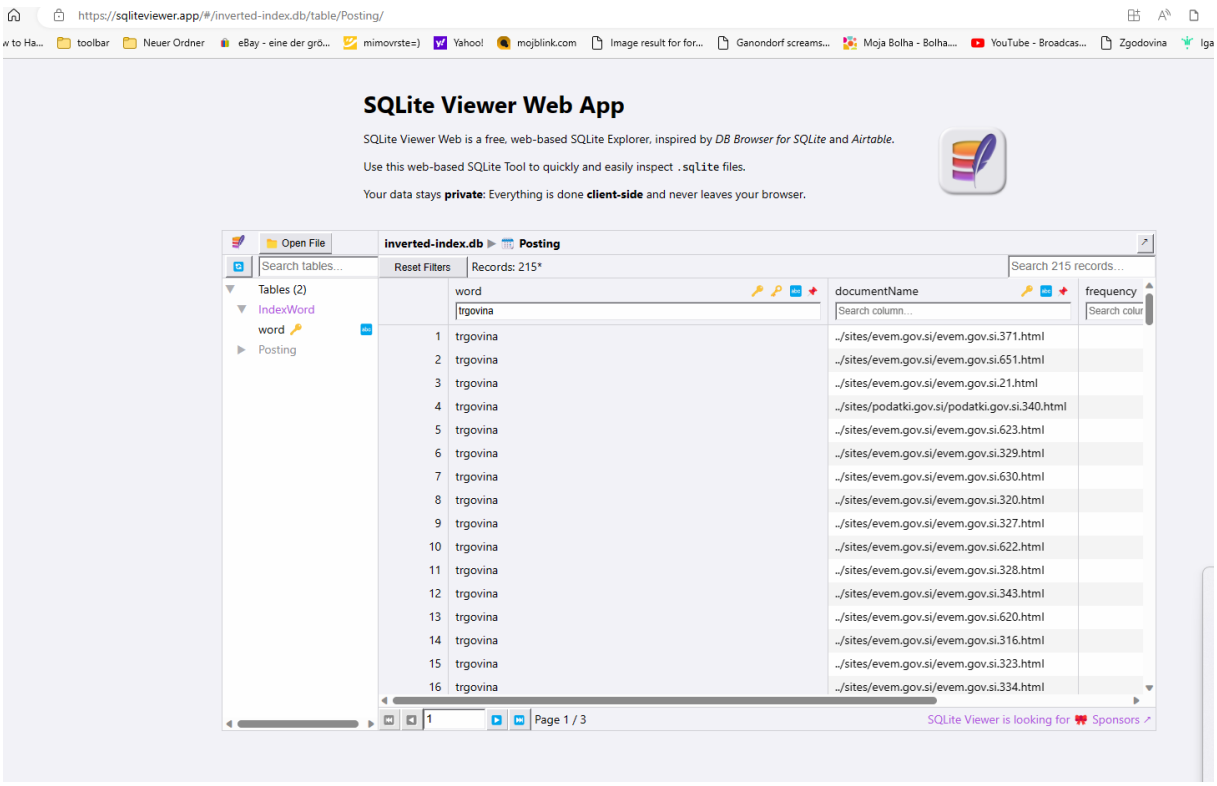
Sestavljena je iz dveh delov. Zunanji del prešteje, za vsako dobljeno stran, število ponovitev vseh besed in združi njihove indekse. V notranjem delu (subquery) pa se za vsako besedo zberejo pojavitve v bazi in se jih združi z ukazom UNION ALL. To naredi naslednja poizvedba:

```
SELECT documentName, frequency, indexes FROM Posting WHERE word = '"+  word
+"'
```

Dobljene rezultate nato problikujemo v primeren izpis.

□ Analiza baze in rezultati

V naši SQLite podatkovni zbirki v tabeli IndexWord se nahaja 35029 zapisov, v tabeli »Posting« pa 352951 zapisov, kjer je se največkrat (upoštevajoč tudi besede, kjer jih izbrana knjižnica za lematizacijo ni uspela popraviti) pojavi beseda »proizvodnja« in sicer 2266-krat samo v dokument evem.gov.si.371.html, ki je hkrati tudi dokument z največ raznolikimi besedami (13301 zapisov), najmanj raznolikimi besedami (31) pa evem.gov.si.55.html. Besede v največ dokumentih so »pogoji« (1398), »uporabe« (1399), »domov« pa 1384. Iskani niz »predelovalne dejavnosti«, se pojavi največrat – 1288 krat v dokumentu evem.gov.si.371.html. Iskani niz »social services« se pojavi najmanjkrat – 12 krat skupno v štirih dokumentih. Za preverbo rezultatov smo uporabili orodje [Posting | Table | inverted-index.db (sqliteviewer.app)](https://sqliteviewer.app/#/inverted-index.db/table/Posting/).



□ Hitrosti poizvedb

Preizkusili smo dve poizvedbi za največkrat pojavljeno besedo “proizvodnja”. S pomočjo inverznega indeks (podatkovne baze), je poizvedba trajala **22 milisekund**, brez indeksa pa je trajala **83 sekund**.

Primer poizvedbe za »avtor franci«.