Indeksiranje in poizvedbe

3. domača naloga pri predmetu WIER

Jakob Merljak, Kristjan Panjan, Blaž Dobravec

MENTOR: asist. prof. dr. Slavko Žitnik

1 Uvod

V tretji seminarski nalogi smo se ukvarjali z implementacijo preprostega iskalnika po html straneh. Naloga je zajemala 3 glavne dele in sicer predprocesiranje in indeksiranje, v katerem vsako spletno stran pretvorimo v ustrezno obliko za nadaljno indekiranje besedila. Druga dva dela se nanašata na poizvedovanja in sicer na poizvedovanje z obrnjenim indeksom in brez obrnjenega indeksa.

2 Predprocesiranje in indeksiranje

Predprocesiranje in indeksiranje se izvede za vsak dokument posebej v več korakih. Za pridobivanje besedila iz HTML dokumentov se uporablja knjižnica Beautiful Soup. Pridobljeno besedilo se nato tokenizira v seznam besed s knjižnico NLTK. Vsako besedo v seznamu se pretvori v male tiskane črke in se jo primerja s seznamom slovenskih štopwordow"knjižnice NLTK, ki je razširjen še z dodatnimi, pri nalogi podanimi štopwordi". Z vsemi besedami, ki se ne ujemajo s štopwordi", se naredi seznam terk, ki vsebujejo besedo in njen odmik (angl. index) v seznamu tokeniziranih besed. Seznam terk besed in njihovih indeksov je rezultat predprocesiranja besedila.

Nato se ta seznam obdela, pri čemer se zgradi slovar besed. Ključ v slovarju so besede, vrednost pa je slovar, ki vsebuje števec pojavitev besede in seznam odmikov besede v tokeniziranem seznamu besed. Pri vsaki ponovitvi besede se števec v slovarju že obstoječe besede poveča in doda nov odmik.

Slovar besed se za tem zapiše v SQLite podatkovno bazo, pri čemer se seznam odmikov pretvori v niz števil, ki so medsebojno ločena z vejico, saj se v bazi odmiki shranjujejo v obliki niza.

3 Poizvedbe z obrnjeni indeksom

Dan poizvedbeni niz se najprej razčleni na seznam besed. Pridobljeni seznam besed se nato predprocesira na enak način kot seznam tokeniziranih besed. Vsako besedo v seznamu se torej pretvori v male tiskane črke in se jih nato primerja s seznamom štopwordov". Vse besede, ki se ne ujemajo s štopwordi", se zbere v nov seznam poizvedbenih besed

S pridobljenimi poizvedbenimi besedami se nato ustvari SQL poizvedba, ki iz tabele *Posting* vrne vse vrstice, kjer se stolpec *word* popolnoma ujema z eno od poizvedbenih besed. SQL poizvedba vrne seznam vrstic z ujemajočimi pojavitvami besed v dokumentih. Ta seznam se nato uporabi za gradnjo slovarja dokumentov. Ključ dokumenta je ime dokumenta, vrednost pa je nov slovar s števcem pojavitev vseh besed poizvedbe v dokumentu, seznamom njihovih odmikov in imenom dokumenta (ime je potrebno v naslednjem koraku, ko bo slovar dokumentov pretvorjen v seznam dokumentov). Slovar dokumentov omogoča, da se pri procesiranju vsake vrnjene vrstice iz baze v konstantnem času posodobi podatke za ustrezen dokument.

Iz slovarja dokumentov se nato pridobi seznam dokumentov (torej vzamemo le vrednosti slovarja), ki se ga sortira glede na skupni števec pojavitev poizvedbenih besed. V seznamu se nato obdrži le konfigurirano število najboljših dokumentov. Za vsak preostali dokument se iz datoteke prebere besedilo s knjižnico Beautiful Soup, besedilo se tokenizira v seznam besed s knjižnico NLTK. Seznam tokeniziranih besed se nato uporabi za gradnjo odsekov (angl. snippet) besedila, kjer se poizvedbene besede nahajajo. Pri tem se uporabi seznam odmikov, ki se je za vsak dokument pridobil iz baze.

Pri grajenju odsekov se najprej iz seznama odmikov tvori seznam terk, kjer vsako terko sestavlja začetni in končni odmik v seznamu tokeniziranih besed, ki sestavljajo odseke. S tem korakom se verižno združuje odseke, kjer se poizvedbena beseda nahaja v odseku prejšnje poizvedbene besede. Za vsako terko se nato začetni odmik zmanjša za nekaj besed in končni odmik poveča za nekaj besed, s čimer se doda primerno okolico iskanim besedam. Število dodanih besed je konfigurabilno. S tema dvema odmikoma se nato iz seznama tokeniziranih besed pridobi rezino seznama (angl. list slice) besed, ki sestavljajo odsek. Rezino se s knjižnico NLTK detokenizira v niz, ki se ga priključi nizu prej pridobljenih odsekov. Niz odsekov se doda v ustrezni slovar dokumenta. Končni rezultat poizvedbe je torej seznam dokumentov s skupnim števcem pojavitev besed in nizom odsekov dokumenta, kjer se nahajajo besede iz poizvedbe.

Ta seznam se nato v ustreznem formatu izpiše na standardni izhod.

4 Poizvedbe brez obrnjenega indeksa

Poizvedbeni niz se procesira na enak način kot pri poizvedbi z indeksom. Za vsak dokument se izvede več korakov. Besedilo dokumenta se predprocesira in procesira na enak način kot pri indeksiranju. Rezultat procesiranja besedila dokumenta je torej slovar besed s številom pojavitev in seznamom odmikov za vsako besedo. Sešteje se

števila pojavitev in združi sezname odmikov vseh besed iz poizvedbe, ki se nahajajo tudi v slovarju. S pridobljenimi vrednostmi se tvori slovar, ki predstavlja dokument, in se ga doda v seznam dokumentov.

Končni rezultat procesiranja vseh dokumentov je seznam dokumentov, ki so predstavljeni s slovarji. Seznam se sortira glede na število pojavitev ter se v njem obdrži konfigurirano število najboljših dokumentov. Seznam se nato v ustreznem formatu izpiše na standardni izhod.

5 Opis baze

(number of indexed words, words and documents with the highest frequencies, . . .) V podatkovni bazi imamo 49082 indeksiranih besed in 402988 ”Posting-ov . V podatkovni bazi nismo odstranjevali dodatnih znakov, ki se pojavljajo v HTML straneh, tako da so v podatkovni bazi vključeni tudi znaki HTML kode. Naredili smo tudi analizo pogostosti posamezne besede (v definiciji besede so tu upoštevani tudi znaki) in ugotovili, da so najbolj frekvenčni znaki, kot so ”,”in ”:”, vendar nas taki znaki v splošnem ne zanimajo. Najbolj frekvenčna zanimiva beseda v enem dokumentu pa je ”nepremičnin”, ki se v dokumentu pojavi 24 krat.

5.1 Poizvedbe

V razdelku preverjanja pravilnosti in ustreznosti rešitev smo obe iskanji pognali na treh primerih.

- edavki

```
> python run-sqlite-search.py "edavki"
Results for a query: "edavki"

Results found in: 0:00:00.166312

Frequencies Document
----------- ----------------------------------------
5           evem.gov.si\evem.gov.si.71.html


3           evem.gov.si\evem.gov.si.74.html

2           evem.gov.si\evem.gov.si.72.html
2           evem.gov.si\evem.gov.si.9.html
1           e-uprava.gov.si\e-uprava.gov.si.32.html

Snippet
----------------------------------------
 ... obliki prek sistema eDavki ali sistema e-VEM ... obliki prek sistema eDavki . Za uporabo ...
 Za uporabo sistema eDavki potrebujete digitalno potrdilo ... DDV-P3 prek sistema eDavki . K zahtevku ...
 Finančna uprava RS eDavki Pomoč & Podpora ...
 ... elektronsko prek portala eDavki najkasneje do 10. ... organ v sistem eDavki ni odložil, ...
 Finančna uprava RS eDavki Pomoč & Podpora ...
 ... obliki prek sistema eDavki . Lestvica za ... Finančna uprava RS eDavki Pravna podlaga Zakon ...
 ... obliki prek sistema eDavki . Za oddajo ... Finančna uprava RS eDavki Pomoč & Podpora ...
 ... . e-Davki Portal eDavki omogoča udobno, ...
```

- predelovalne dejavnosti

```
> python run-sqlite-search.py "predelovalne dejavnosti"
Results for a query: "predelovalne dejavnosti"

Results found in: 0:00:05.095892

Frequencies Document
----------- ----------------------------------------
1288        evem.gov.si\evem.gov.si.371.html
75          evem.gov.si\evem.gov.si.377.html
40          podatki.gov.si\podatki.gov.si.340.html
39          evem.gov.si\evem.gov.si.452.html
31          evem.gov.si\evem.gov.si.653.html

Snippet
----------------------------------------
 ... iskanje ustrezne šifre dejavnosti /storitve in informacij ... pogojih za opravljanje dejavnosti . V iskalnik ... 845 od 9
 ... Defektolog v zdravstveni dejavnosti Dekan oziroma direktor ... Dietetik v zdravstveni dejavnosti Dimikar Diplomirana med
 ... - NOSILEC DOPOLNILNE DEJAVNOSTI NA KMETIJI BREGAR ... šport CENTER INTERESNIH DEJAVNOSTI PTUJ CENTER JUDOVSKE ... ŠOLSKI
 ... Druge storitvene dejavnosti, drugje nerazvrščene ... 96.090) / Dejavnosti / eVEM Republika ... e-VEM eVEM>Dejavnosti>Drug
 ... Dovoljenje za opravljanje dejavnosti specializirane prodajalne z ... radijske ali televizijske dejavnosti Dovoljenje za
```

- social services

```
> python run-sqlite-search.py "social services"
Results for a query: "social services"

Results found in: 0:00:01.603678

Frequencies Document
----------- ----------------------------------------
5           e-uprava.gov.si\e-uprava.gov.si.45.html
5           e-uprava.gov.si\e-uprava.gov.si.9.html
1           evem.gov.si\evem.gov.si.661.html
1           podatki.gov.si\podatki.gov.si.340.html

Snippet
----------------------------------------
 ... , retirement Social services, health, ... etc.? Social services, health, ... Labour,
 ... , retirement Social services, health, ... etc.? Social services, health, ... Labour,
 ... Records and Related Services (AJPES) ...
 ... recreation and spa services ltd. TERME MARIBOR ...
```

- trgovina

```
> python run-sqlite-search.py "trgovina"
Results for a query: "trgovina"

Results found in: 0:00:05.033538

Frequencies Document
----------- ----------------------------------------
364         evem.gov.si\evem.gov.si.371.html

94          evem.gov.si\evem.gov.si.651.html
92          evem.gov.si\evem.gov.si.21.html
82          podatki.gov.si\podatki.gov.si.340.html
13          evem.gov.si\evem.gov.si.623.html

Snippet
----------------------------------------
 ... gl . 46.110 trgovina na debelo s ... gl . 10.890 trgovina na debelo z ... gl . 10.890 trgovina na debelo s ... gl . 4
 ... Skladiščenje nevarnih kemikalij Trgovina na debelo z ... z nevarnimi kemikalijami Trgovina na drobno z ... gl . 32.50
 ... Druga govedoreja Druga trgovina na drobno v ... specializiranih prodajalnah Druga trgovina na drobno v ... nespecializi
 ... Moj e-VEM eVEM>Področja Trgovina Tu boste našli ... Seznam dejavnosti Druga trgovina na drobno v ... nespecializiranih
 ... A DENT, trgovina in storitve, ... . ADRIA INVESTICIJE trgovina, posredništvo, ... d.o.o . AHATSERVIS trgovina in stor
 ... Trgovina na debelo z ... izdelki široke porabe Trgovina na debelo z ... Sem spada: trgovina na debelo z ... izdelki i
```

- dohodninska olajšava

```
> python run-sqlite-search.py "dohodninska olajšava"
Results for a query: "dohodninska olajšava"
Results found in: 0:00:00.189542
Frequencies Document
----------- ----------------------------------------
1           evem.gov.si\evem.gov.si.22.html
1           evem.gov.si\evem.gov.si.77.html
1           podatki.gov.si\podatki.gov.si.211.html

Snippet
----------------------------------------
 ... 50% (dohodninska lestvica) Davek ...
 ... upoštevana le splošna olajšava, pri normiranih ...
 ... , obresti, olajšava, plača, ...
```

Prvi primer pokaže delovanje algoritma, ki prioritizira vire podatkov, v katerih je frekvenca iskanega niza največja velja pa tudi, da je ta stran najbolj ustrezna, če nas zanimajo edavki. V drugem primeru se za besedno zvezo ”predelovalne dejavnosti”v prvem dokumentu pojavi kar 1288 zadetkov. V tretjem pa je možno opaziti, da iskanje res izvaja razbitje besedne zveze šocial services”, saj iskalnik pokaže tudi primer, ki se nanaša na ”recreation and spa **services**”. Dodatno smo izbrali še ”dohodninska olajšava”.

6 Zaključek

Implementirane so bile vse točke domače naloge, analizirli pa smo tudi različne poizvedbe ter preverili ustreznost primera ko uporabljamo obrnjen indeks in ko le ta ni obrnjen. Metoda z obrnjenim algoritmom je bila v povprečju hitrejša približno za faktor 60 (v primeru, da je metoda z obrnjenim indeksom potrebovala 1 sekundo je metoda brez obrnjenega indeks potrebovala 1 minuto).