Ekstrakcija podatkov iz spleta

3. domača naloga pri predmetu IEPS

Žiga Černigoj 63130028

Marko Lavrinec 63130134

Maj 2019

1 Uvod

Cilj naloge je razviti program, ki za podan iskalni niz poišče ustrezne rezultate v bazi in datotekah. Za implementacijo bo uporabljen programski jezik Python.

2 Pred-procesiranje in indeksiranje podatkov

Skripta za pred-procesiranje najprej preveri, če že obstaja SQLite datoteka in jo v nasprotnem primeru naredi s pravilno zasnovo tabel.

Nato se skripta sprehodi čez vse datoteke v imeniku. Iz vsake datoteke z uporabo knjižnice BeautifulSoup prebere vsebino znotraj značke `body` in pri tem odstrani značke `script`, `noscript` in `style`, katerih vsebina bi se drugače prav tako uporabila pri nadaljnjih korakih. Skripta nato izvede tokenizacijo - ustvari tabelo besed in ločil, ki so razporejeni v takem vrstnem redu, kot so v besedilu. Za ustvarjanje `postingov` skripta uporabi le tokene (žetone - posamezne besede), ki niso *stopword-i* in ločila, pred uporabo pa tokene pretvori v enak format - male črke.

Za tem skripta tokene, ki zadostujejo pogojem (t.j. so besede oziroma niso *stopword-i* ali ločila), vstavi v slovar in zabeleži vse njihove pojavitve ter število pojavitev. Za mesto pojavitve se zabeleži zaporedni indeks v izvirni tabeli tokenov.

Prvotni način vstavljanja podatkov v bazo je bil tak, da smo vsako besedo iz slovarja najprej vstavili v tabelo *IndexWord*, če ta v tabeli še ne obstaja, nato pa še v tabelo *Posting*, kamor smo shranili še ime dokumenta kjer se nahaja, frekvenco ter indekse pojavitev, ločene z vejicami. Pri tem smo vsak vnos posebej tudi potrdili (*commit-ali*).

Nato smo se odločili za uporabo transakcij z več vnosi naenkrat. Tako najprej v bazo v tabelo *IndexWord* vstavimo vse besede iz vseh dokumentov in nato naredimo potrditev (*commit*). Nato za vsak dokument vstavimo vse postinge v tabelo *Posting* in šele nato naredimo potrditev (*commit*). Tak način se je izkazal za veliko izboljšavo, saj je namreč pohitril shranjevanje podatkov v bazo za okrog 8x.

3 Iskanje po bazi in Inverted index

Za podani iskalni niz se ta najprej tokenizira in nato v bazi poiščejo vse pojavitve teh besed z naslednjim SQL stavkom:

```
SELECT word, documentName, frequency, indexes
FROM Posting
WHERE word IN ({seq})
```

Dobljeni rezultat se združi glede na dobljene dokumente in izračuna ustreznost in frekvenca pojavitve.

Dobljeni rezultati se nato v zaporednem vrstnem redu izpišejo uporabniku skladno z navodili naloge.

4 Zaporedno branje datotek

Skripta se sprehodi čez vse datoteke, jih tokenizira in nato v njej poišče vse pojavitve iskanega niza.

Samo delovanje takega iskanja je počasno, saj se vsakič na novo procesira vsebina vseh datotek in po njih išče ustrezne besede.

5 Rezultati in analiza

Za sledeče nize smo dobili rezultate:

- "predelovalne dejavnosti"

  Zaradi preobsežnega rezultata je ostalo na voljo v mapi results/predelovalne_dejavnosti.txt. Prav tako je snippet omejen na 200 znakov.

  Results for a query: " predelovalne dejavnosti "

  Results found in 0.015369892120361328

  Frequencies Document Snippet

  1291 ./PA3-data/evem.gov.si/evem.gov.si.371.html ...iskanje ustrezne šifre dejavnosti /storitve in informacij ... ...pogojih za opravljanje dejavnosti . V iskalnik ... ...645 od 645 dejavnosti Izpisanih je od ... ...Izpisanih je od dejavnosti A KMET

  75 ./PA3-data/evem.gov.si/evem.gov.si.377.html ...Defektolog v zdravstveni dejavnosti Dekan oziroma direktor ... ...Dietetik v zdravstveni dejavnosti Dimnikar Diplomirana medicinska ... ...I v zdravstveni dejavnosti Laboratorijski sodelavec II ...

40 ./PA3-data/evem.gov.si/evem.gov.si.452.html ...nastavitve Druge storitvene dejavnosti , drugje nerazvrščene ... ...96.090) / Dejavnosti / eVEM Republika ... ...e-VEM eVEM Dejavnosti Druge storitvene ... ... Druge storitvene dejavnosti , d

40 ./PA3-data/podatki.gov.si/podatki.gov.si.340.html ...- NOSILEC DOPOLNILNE DEJAVNOSTI NA KMETIJI BREGAR ... ...šport CENTER INTERESNIH DEJAVNOSTI PTUJ CENTER JUDOVSKE ... ...ŠOLSKIH IN OBŠOLSKIH DEJAVNOSTI Center urbane kulture ... ...in druge zdravst

- "trgovina"

  Zaradi preobsežnega rezultata je ostalo na voljo v mapi results/trgoina.txt. Prav tako je snippet omejen na 200 znakov.

  esults for a query: " trgovina "

  Results found in 0.03357744216918945

  Frequencies Document Snippet

  ---

  728 ./PA3-data/evem.gov.si/evem.gov.si.371.html ...gl . 46.110 trgovina na debelo s ... ...gl . 46.110 trgovina na debelo s ... ...gl . 10.890 trgovina na debelo z ... ...gl . 10.890 trgovina na debelo z ... ...gl . 10.890 trgovina na debelo s ...

  192 ./PA3-data/evem.gov.si/evem.gov.si.651.html ...Druga govedoreja Druga trgovina na drobno v ... ...Druga govedoreja Druga trgovina na drobno v ... ...specializiranih prodajalnah Druga trgovina na drobno v ... ...specializiranih prodajalnah Druga

  184 ./PA3-data/evem.gov.si/evem.gov.si.21.html ...eVEM Področja Trgovina Tu boste našli ... ...eVEM Področja Trgovina Tu boste našli ... ...Seznam dejavnosti Druga trgovina na drobno v ... ...Seznam dejavnosti Druga trgovina na drobno v ... ..

  164 ./PA3-data/podatki.gov.si/podatki.gov.si.340.html ...A DENT , trgovina in storitve , ...A DENT , trgovina in storitve , ... ADRIA INVESTICIJE trgovina , posredništvo , ... ADRIA INVESTICIJE trgovina , posredništvo , ...d.o.o . AHATS

- "social services"

  Results for a query: " social services "

  Results found in 0.006722927093505859

  Frequencies Document Snippet

  ---

4 ./PA3-data/e-uprava.gov.si/e-uprava.gov.si.45.html ...Labour , retirement Social services , health , ...relationship etc. ? Social services , health , ...I obtain financial social assistance ? How ...

4 ./PA3-data/e-uprava.gov.si/e-uprava.gov.si.9.html ...Labour , retirement Social services , health , ...relationship etc. ? Social services , health , ...I obtain financial social assistance ? How ...

1 ./PA3-data/evem.gov.si/evem.gov.si.661.html ...Records and Related Services (AJPES) ...

1 ./PA3-data/podatki.gov.si/podatki.gov.si.340.html ...recreation and spa services ltd. TERME MARIBOR ...

-
-
-

6 Zagon algoritma

Zagon programa za indeksiranje se lahko izvede preko ukazne vrstice z vpisom ukaza `python data_processor.py` , za iskanje rezultatov z uporabo invernega indeksa je potrebno pognati `python index_search.py` , za pridobitev rezultatov neposredno iz datotek (na "star", počasen način) pa `python naive_search.py` .

7 Zaključek

V tej seminarski nalogi smo razvili delujoči program za indeksiranje vsebin spletnih strani.

References