# Poročilo 3. seminarske naloge

Robert Košir, Aljoša Omejc

May 22, 2020

# 1 Uvod

V naslednjih nekaj poglavjih bomo na kratko opisali projekt pri predmetu Iskanje in ekstrakcija podatkov s spleta. Sam projekt se ukvarja generiranjem indeksa z bazo besed na 1416 straneh gov.si domene. Po generiranju pa se lahko ta indeks uporabi za iskanje ključnih besed. Prav tako, pa se lahko išče ključne besede brez predhodnje baze, vendar ta postopek traja nekoliko dlje.

# 2 Implementacija

## 2.1 Indeksiranje dokumentov

Za učinkovito iskanje najprej potrebujemo bazo s podatki o besedah. Te smo dobili tako, da smo se sprehodili čez vse podane dokumente prebrali njihovo telo (*ang. bofy*) in odstranili odvečne stvari:

- script
- style
- link
- iframe
- noscript

Tako smo se znebili besed, ki so zapisani v zgoraj omenjenih vrsticah teksta. Nato smo pridobili tekst z uporabo *get_text()* funkcije in jo celotno poenostavili na majhne črke. Nato pa smo tekst poslali v metodo *process(tekst, url)*, ki sprejme očiščen test in pa stran, katero preiskujemo. Z uporabo metode *word_tokenize()* smo pridobili vse besede in iz njih izločili tako imenovane "stop besede". Poiskali smo še indekse vsake besede in jih shranili v bazo.

## 2.2 Iskanje po dokumentih z že zgrajenim indekson

Vhodni parameter najprej tokeniziramo in iz njega pridobimo ključne besede. Poiščemo vse pojavitve vhodnih besed in jih združimo po dokumentih. Indekse besed, ki so si zelo blizu (20 znakov) združimo v isti delček (angl. snippet). Ponovno odpremo vse dokumente, kjer se nahajajo ključne besede, jih s procesiramo in izpišemo okolico besed. Zaradi ponovnega branja datotek in procesiranja besedila, lahko to traja tudi do nekaj sekund. Izboljšave za iskanje okolice, bi naredila tako, da sprocesirano besedilo shraniva že kar v bazo, kar bi znatno pohitrilo iskanje.

## 2.3 Iskanje po dokumentih brez zgrajenega indeksa

Ta del je zelo podoben zgornjemu, le da moramo namesto iskanja po bazi, generirati indeks vsakič znova. To nam poslabša čas iskanja. Dokumente pregledamo po istem postopku kot v poglavju 2.1 le da ne iščemo indeksa vsake besede, ampak si shranimo le v kateremu dokumenta se ta beseda nahaja. Nato pa podatke pošljemo v enako funkcijo kot pri poglavju 2.2.

# 3 Rezultati

## 3.1 Opis baze

Baza ima 49114 indeksiranih besed. Največkrat se pojavijo ločila. Besede, ki pa so najvišje pa so:

- proizvodnja - 2266 pojavitev
- gl - 1668 pojavitev
- spada - 1338 pojavitev
- dejavnosti - 1284 pojavitev

Največ najdenih besed vsebujejo naslednje strani:

- evem.gov.si.371.html - 13195 besed
- podatki.gov.si.340.html - 6568 besed
- e-prostor.gov.si.166.html - 6102 besed

Najmanj pa *evem.gov.si.55.html*, ki vsebuje le 20 besed.

## 3.2 Predstavitev iskanih besed

Iz spodnjih slik je razvidno, da program dlje časa obdeluje podatke, ko je večje število dokumentov, ki vsebuje to datoteko, saj mora program vsak dokument odpreti in tako izgubimo na hitrosti. Kot sva omenila že zgoraj bi lahko namesto odpiranja dokumentov besedilo direktno shranjevala v bazo.
Brez zgrajene indeksa pa program potrebuje približno 90-110 sekund, da poišče želeno besedo.

```
Results for a query:predelovalne dejavnosti
    Results found in  21.350057363510132
Frequenices    Document                                        Snippet
-------------  ----------------------------------------------  ------------------------------------------------------------------------------------------------------
1570           evem.gov.si/evem.gov.si.371.html                ...anje ustrezne šifre dejavnosti /storitve......ojih za opravljanje dejavnosti.  v iskal......sanih je 645 od 645 dejavnosti
78             evem.gov.si/evem.gov.si.377.html                ...tolog v zdravstveni dejavnosti  dekan oz......tetik v zdravstveni dejavnosti  dimnikar..... delavci v sevalnih dejavnostih
47             evem.gov.si/evem.gov.si.452.html                ...oj e-vem       evem›dejavnosti›druge sto.....ti›druge storitvene dejavnosti, drugje n.....   druge storitvene dejavnosti,
40             podatki.gov.si/podatki.gov.si.340.html          ... nosilec dopolnilne dejavnosti na kmetij.....  center interesnih dejavnosti ptuj     .....olskih in obšolskih dejavnosti
31             evem.gov.si/evem.gov.si.398.html                ...jene na opravljanje dejavnosti (npr.: pr..... namene opravljanja dejavnosti ipd.   v .....ecev z opravljanjem dejavnosti
31             evem.gov.si/evem.gov.si.653.html                ...enje za opravljanje dejavnosti specializ.....ke ali televizijske dejavnosti  dovoljen.....a izvajanje sevalne dejavnosti
29             evem.gov.si/evem.gov.si.72.html                 ...davek od dohodka iz dejavnosti  davek od.....davek od dohodka iz dejavnosti     ko za.....davek od dohodka iz dejavnosti.
28             evem.gov.si/evem.gov.si.442.html                ...oj e-vem       evem›dejavnosti›dejavnost.....    evem›dejavnosti›dejavnosti za nego t.....lesa (96.040)       dejavnosti
22             evem.gov.si/evem.gov.si.265.html                ...oj e-vem       evem›dejavnosti›proizvodn.....ta skd šifra zajema dejavnosti in storit.....etek in opravljanje dejavnosti.
22             evem.gov.si/evem.gov.si.276.html                ...oj e-vem       evem›dejavnosti›storitve .....ta skd šifra zajema dejavnosti in storit.....etek in opravljanje dejavnosti.
```

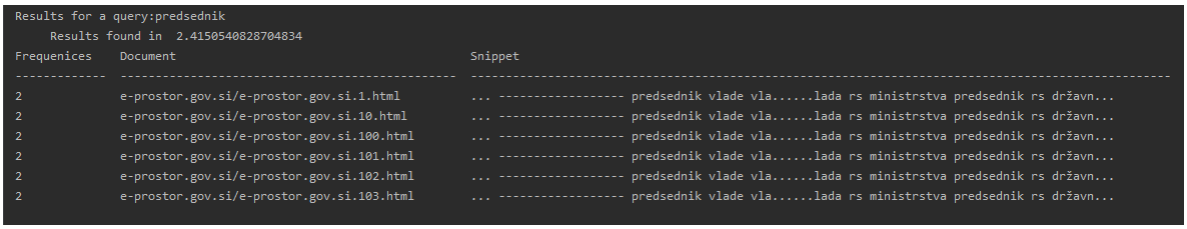
Figure 1: Predelovalne dejavnosti

```
Results for a query:trgovina
    Results found in  6.195161581039429
Frequenices    Document                                        Snippet
-------------  ----------------------------------------------  ------------------------------------------------------------------------------------------------------
368            evem.gov.si/evem.gov.si.371.html                ...nizacij, gl. 46.110 trgovina na debelo s..... in juh, gl. 10.890 trgovina na debelo z.....ov ipd., gl. 10.890
96             evem.gov.si/evem.gov.si.651.html                ...a govedoreja  druga trgovina na drobno v..... prodajalnah  druga trgovina na drobno v..... prodajalnah  druga
92             evem.gov.si/evem.gov.si.21.html                 ...evem›področja       trgovina  tu boste n.....m dejavnosti  druga trgovina na drobno v..... prodajalnah  druga
82             podatki.gov.si/podatki.gov.si.340.html          ...           a dent, trgovina in storitve.....  adria investicije trgovina, posredništ.....         ahatservis
13             evem.gov.si/evem.gov.si.623.html                ...    evem›dejavnosti›trgovina na debelo z.....široke porabe       trgovina na debelo z.....porabe  sem spada:
12             evem.gov.si/evem.gov.si.329.html                ...    evem›dejavnosti›trgovina na debelo z.....itarno opremo       trgovina na debelo z.....premo  sem spada:
```

Figure 2: Trgovina

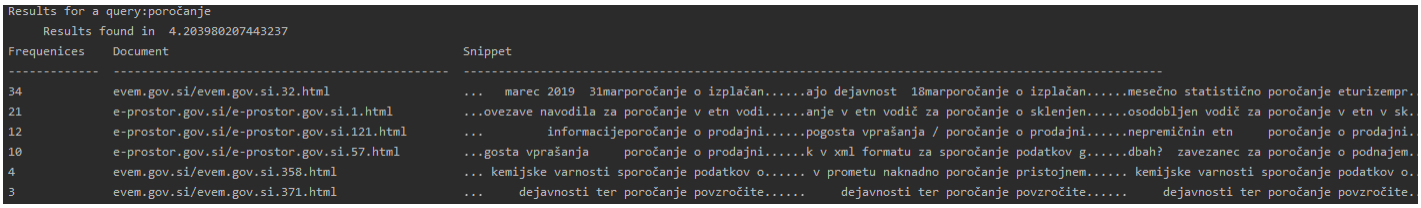
```
Results for a query:social services
    Results found in  1.18062162399292
Frequenices    Document                                        Snippet
-------------  ----------------------------------------------  ------------------------------------------------------------------------------------------------------
5              e-uprava.gov.si/e-uprava.gov.si.45.html         ...abour, retirement   social services, hea.....ationship etc.?     social services, hea..... i obtain financial social assistance? h...
5              e-uprava.gov.si/e-uprava.gov.si.9.html          ...abour, retirement   social services, hea.....ationship etc.?     social services, hea..... i obtain financial social assistance? h...
1              evem.gov.si/evem.gov.si.661.html                ...records and related services (ajpes) and...
1              podatki.gov.si/podatki.gov.si.340.html          ... recreation and spa services ltd.       ...
```

Figure 3: Social services



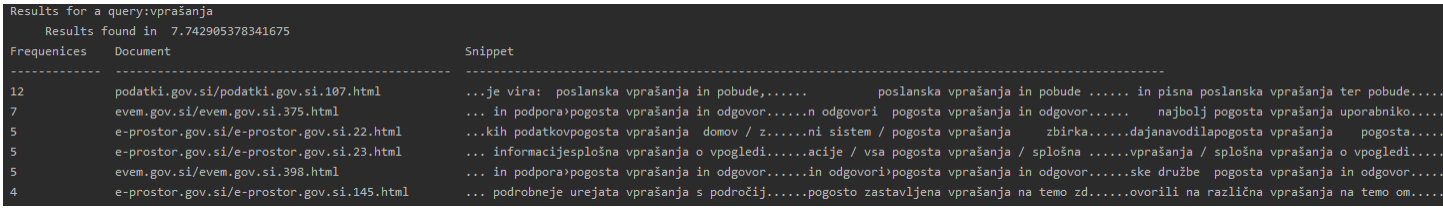
```
Results for a query:predsednik
     Results found in  2.4150540828704834
Frequenices    Document                                        Snippet
-------------  ----------------------------------------------  ----------------------------------------------------------------------------------------------------
2              e-prostor.gov.si/e-prostor.gov.si.1.html        ... ------------------ predsednik vlade vla......lada rs ministrstva predsednik rs državn...
2              e-prostor.gov.si/e-prostor.gov.si.10.html       ... ------------------ predsednik vlade vla......lada rs ministrstva predsednik rs državn...
2              e-prostor.gov.si/e-prostor.gov.si.100.html      ... ------------------ predsednik vlade vla......lada rs ministrstva predsednik rs državn...
2              e-prostor.gov.si/e-prostor.gov.si.101.html      ... ------------------ predsednik vlade vla......lada rs ministrstva predsednik rs državn...
2              e-prostor.gov.si/e-prostor.gov.si.102.html      ... ------------------ predsednik vlade vla......lada rs ministrstva predsednik rs državn...
2              e-prostor.gov.si/e-prostor.gov.si.103.html      ... ------------------ predsednik vlade vla......lada rs ministrstva predsednik rs državn...
```

Figure 4: Predsednik



```
Results for a query:poročanje
     Results found in  4.203980207443237
Frequenices    Document                                        Snippet
-------------  ----------------------------------------------  ----------------------------------------------------------------------------------------------------
34             evem.gov.si/evem.gov.si.32.html                 ...   marec 2019  31marporočanje o izplačan......ajo dejavnost  18marporočanje o izplačan......mesečno statistično poročanje eturizempr.
21             e-prostor.gov.si/e-prostor.gov.si.1.html        ...ovezave navodila za poročanje v etn vodi......anje v etn vodič za poročanje o sklenjen......osodobljen vodič za poročanje v etn v sk.
12             e-prostor.gov.si/e-prostor.gov.si.121.html      ...        informacijeporočanje o prodajni......pogosta vprašanja / poročanje o prodajni......nepremičnin etn     poročanje o prodajni.
10             e-prostor.gov.si/e-prostor.gov.si.57.html       ...gosta vprašanja     poročanje o prodajni......k v xml formatu za sporočanje podatkov g......dbah?  zavezanec za poročanje o podnajem.
4              evem.gov.si/evem.gov.si.358.html                ... kemijske varnosti sporočanje podatkov o..... v prometu naknadno poročanje pristojnem..... kemijske varnosti sporočanje podatkov o.
3              evem.gov.si/evem.gov.si.371.html                ...     dejavnosti ter poročanje povzročite.....     dejavnosti ter poročanje povzročite.....     dejavnosti ter poročanje povzročite.
```

Figure 5: Porocanje



```
Results for a query:vprašanja
     Results found in  7.742905378341675
Frequenices    Document                                        Snippet
-------------  ----------------------------------------------  ----------------------------------------------------------------------------------------------------
12             podatki.gov.si/podatki.gov.si.107.html          ...je vira:  poslanska vprašanja in pobude,......          poslanska vprašanja in pobude ..... in pisna poslanska vprašanja ter pobude.....
7              evem.gov.si/evem.gov.si.375.html                ... in podpora›pogosta vprašanja in odgovor......n odgovori  pogosta vprašanja in odgovor.....    najbolj pogosta vprašanja uporabniko.....
5              e-prostor.gov.si/e-prostor.gov.si.22.html       ...kih podatkovpogosta vprašanja  domov / z......ni sistem / pogosta vprašanja     zbirka......dajanavodilapogosta vprašanja    pogosta.....
5              e-prostor.gov.si/e-prostor.gov.si.23.html       ... informacijesplošna vprašanja o vpogledi......acije / vsa pogosta vprašanja / splošna ......vprašanja / splošna vprašanja o vpogledi.....
5              evem.gov.si/evem.gov.si.398.html                ... in podpora›pogosta vprašanja in odgovor......in odgovori›pogosta vprašanja in odgovor......ske družbe  pogosta vprašanja in odgovor.....
4              e-prostor.gov.si/e-prostor.gov.si.145.html      ... podrobneje urejata vprašanja s področij......pogosto zastavljena vprašanja na temo zd......ovorili na različna vprašanja na temo om.....
```

Figure 6: Vprasanja