IEPS Poročilo 3 - Ekipa “Vipavska Burja”

Matjaž Rupnik, Blaž Ličen

Maj 2021

1 Specifike in odločitve

1.1 Data processing with indexing

Indeksiranje podatkov oz. besed smo realizirali s pomočjo knjižnic”BeautifulSoup”in nltk ”word_tokenize”. Najprej smo za vsako stran (datoteko) prebrali njeno vsebino, jo razdelili na tokene, jih formatirali v zapis z malimi črkami (lowercase), odstranili stopworde (ki smo jih dopolnili z nekaj ločili) in odstranili nekaj števil in povezav. Tako je ostal le seznam besed, ki so pripravljene za zapis v bazo. Za vsako datoteko smo najprej v tabelo IndexWord shranili morebitne nove besede, nato pa smo v tabelo Posting shranili frekvence in indekse pojavitev besed v besedilu. Prav tako pa smo tekste vseh datotek shranili v pickle, ki je bil nato uporabljen za prikaz snippeta. Tako nam ni bilo treba ob poizvedbi odpirati datotek.

1.2 Data retrieval with inverted index

Poizvedbe z invertiranim indekson delujejo tako, da najprej gremo v bazo, ter izpišemo vse datoteke v katerih se beseda nahaja, ter indekse. Nato se pripravi snippet, za to se uporabi pickle, ki se je pripravil ob grajenju podatkovne baze. Zatem pa se rezultati izpišejo na zaslon. Take poizvedbe navadno trajajo zelo malo časa, saj je potrebna samo ena poizvedba v podatkovni bazi ter izpis rezultatov.

1.3 Data retrieval without inverted index

Poizvedbe z iskanjem po datotekah delujejo tako, da gremo z zanko skozi vse datoteke, ter iz vsake datoteke pridobimo besedilo, ter nato gremo z zanko skozi vse besede, ter gledamo ali se beseda ujema z besedo iz pozvedbe in če hkrati ni v štop wordih”. Če se iskana beseda ujema s trenutno besedo, potem lahko to besedo in njeno okolico zapišemo v snippet, ki se na koncu izpiše na zaslon.

2 Podatkovna baza

Število zapisov	Tabela v bazi
40717	IndexWord
363991	Posting

Tabela 1: Število vseh zapisov v bazi

	Število različnih besed	Stran
1	12431	evem.gov.si\evem.gov.si.371.html
2	6535	podatki.gov.si\podatki.gov.si.340.html
3	3433	e-prostor.gov.si\e-prostor.gov.si.166.html
4	2461	e-prostor.gov.si\e-prostor.gov.si.218.html
5	1642	e-prostor.gov.si\e-prostor.gov.si.57.html
6	1511	evem.gov.si\evem.gov.si.398.html
7	1268	evem.gov.si\evem.gov.si.651.html
8	1164	e-uprava.gov.si\e-uprava.gov.si.56.html
9	1010	evem.gov.si\evem.gov.si.653.html
10	976	e-uprava.gov.si\e-uprava.gov.si.44.html

Tabela 2: Deset strani z največ različnimi besedami

	Število besed	Stran
1	78397	evem.gov.si\evem.gov.si.371.html
2	27358	podatki.gov.si\podatki.gov.si.340.html
3	6029	e-prostor.gov.si\e-prostor.gov.si.166.html
4	4084	evem.gov.si\evem.gov.si.398.html
5	3533	e-prostor.gov.si\e-prostor.gov.si.147.html
6	3503	e-prostor.gov.si\e-prostor.gov.si.57.html
7	3434	podatki.gov.si\podatki.gov.si.511.html
8	2867	e-prostor.gov.si\e-prostor.gov.si.150.html
9	2595	evem.gov.si\evem.gov.si.651.html
10	2498	evem.gov.si\evem.gov.si.653.html

Tabela 3: Deset strani z največ besedami

3 Rezultati testnih poizvedb

Rezultati dejanskih poizvedb so prikazani v spodnjih podpoglavjih, jasen pa je trend. Poizvedbe, ki so uporabljale podatkovno bazo za indeksiranje trajajo v rangu nekaj milisekund, medtem ko poizvedbe, ki ne uporabljajo baze, ampak sproti iščejo besede po besedilu dokumenta, trajajo nekje minuto, odvisno od zmogljivosti računalnika.

3.0.1 Poizvedba 1: predelovalne dejavnosti

Poizvedba vrne 2060 rezultatov.

```
(Nova mapa) C:\Users\Toncaw\Documents\IEPS\1. projekt\ieps-vipavska-burja\pa3\implementation-indexing>run-sqlite-search.py predelovalne dejavnosti

Results for a query: "predelovalne dejavnosti"

        Results found in 13.142 ms

        Frequencies Document                                Snippet
        ----------- --------------------------------------- ------------------------------------------------------------
        1291        evem.gov.si\evem.gov.si.371.html           ... iskanje ustrezne šifre dejavnosti /storitve in informacij o pogojih za opravljanj
 KMETIJSTVO IN ... pogojih za opravljanje dejavnosti: · Pridelava semenskega ... pogojih za opravljanje dejavnosti: · Pridelava kmetijskih ... pogoj
 dejavnosti: Pridelava semenskega materiala ... pogojih za opravljanje dejavnosti: Pridelava in distribucija ... pogojih za opravljanje dejavnosti:
kih pridelkov ... pogojih za opravljanje dejavnosti: Pridelava semenskega materiala ... Pogoji: · Lista dejavnosti, ki se običajno ... obrtni način
e dejavnosti poslovni subjekti po ... pogojih za opravljanje dejavnosti: Pridelava semenskega materiala ... pogojih za opravljanje dejavnosti: Pride
kov ... pogojih za opravljanje dejavnosti: Pridelava kmetijskih pridelkov ... pogojih za opravljanje dejavnosti: Pridelava kmetijskih pridelkov ...
javnosti: Pridelava kmetijskih pridelkov ... pogojih za opravljanje dejavnosti: Pridelava kmetijskih pridelkov ... pogojih za opravljanje dejavnosti
```

Slika 1: predelovalne dejavnosti - sql search

```
(Nova mapa) C:\Users\Toncaw\Documents\IEPS\1. projekt\ieps-vipavska-burja\pa3\implementation-indexing>run-basic-search.py predelovalne dejavnosti

Results for a query: "predelovalne dejavnosti"

        Results found in 76.961 s

        Frequencies Document                                Snippet
        ----------- --------------------------------------- ------------------------------------------------------------
        1291        evem.gov.si.371.html                    ... iskanje ustrezne šifre dejavnosti /storitve in informacij o... pogojih za opravljan
A KMETIJSTVO IN LOV... pogojih za opravljanje dejavnosti: · Pridelava semenskega... pogojih za opravljanje dejavnosti: · Pridelava kmetijskih... pogo
e dejavnosti: Pridelava semenskega materiala... pogojih za opravljanje dejavnosti: Pridelava in distribucija... pogojih za opravljanje dejavnosti: Pr
h pridelkov... pogojih za opravljanje dejavnosti: Pridelava semenskega materiala... aster) Pogoji: · Lista dejavnosti, ki se običajno... obrtni nači
egistraciji te dejavnosti poslovni subjekti po uradni... pogojih za opravljanje dejavnosti: Pridelava semenskega materiala... pogojih za opravljanje
kmetijskih pridelkov... pogojih za opravljanje dejavnosti: Pridelava kmetijskih pridelkov... pogojih za opravljanje dejavnosti: Pridelava kmetijskih
pravljanje dejavnosti: Pridelava kmetijskih pridelkov... pogojih za opravljanje dejavnosti: Pridelava kmetijskih pridelkov... pogojih za opravljanje
```

Slika 2: predelovalne dejavnosti - basic search

3.0.2 Poizvedba 2: trgovina

Poizvedba vrne 347 rezultatov.

```
(Nova mapa) C:\Users\Toncaw\Documents\IEPS\1. projekt\ieps-vipavska-burja\pa3\implementation-indexing>run-sqlite-search.py trgovina

Results for a query: "trgovina"

        Results found in 18.998 ms

        Frequencies Document                                Snippet
        ----------- --------------------------------------- ------------------------------------------------------------
        364         evem.gov.si\evem.gov.si.371.html           ... organizacij, gl. 46.110 trgovina na debelo s ... juh, gl. 10.890 trgovina n
 s ... Skladiščenje nevarnih kemikalij Trgovina na debelo z nevarnimi kemikalijami Trgovina na drobno z ... kompres, gl. 32.500 trgovina na de
 in prodaja ... odpadki, gl. 38.220 trgovina na debelo z ... in tehnologijo G TRGOVINA; VZDRŽEVANJE IN POPRAVILA MOTORNIH VOZIL 45 Trgovina z
ovina na debelo in ... Sem ne spada: trgovina na debelo ali ... na drobno 45.190 Trgovina z drugimi motornimi vozili Sem spada: trgovina na de
z ... vozila Sem spada: trgovina na debelo s ... motorna vozila 45.320 Trgovina na drobno z ... vozila Sem spada: trgovina na drobno s ... Sem
 motornih koles; trgovina z njihovimi deli ... opremo Sem spada: trgovina na debelo ali ... tudi z mopedi trgovina na debelo ali ... za motorn
.490 trgovina na drobno s ... 46 Posredništvo in trgovina na debelo, razen ... Sem ne spada: trgovina na debelo za ... Sem ne spada: trgovina
```

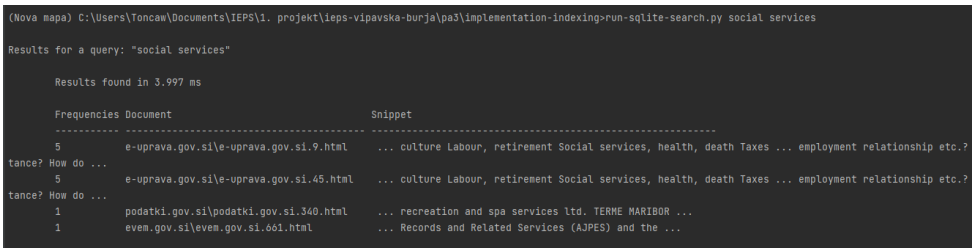
Slika 3: trgovina - sql search

```
(Nova mapa) C:\Users\Toncaw\Documents\IEPS\1. projekt\ieps-vipavska-burja\pa3\implementation-indexing>run-basic-search.py trgovina

Results for a query: "trgovina"

        Results found in 64.019 s

        Frequencies Document                                Snippet
        ----------- --------------------------------------- ------------------------------------------------------------
        364         evem.gov.si.371.html                    ... zadružnih organizacij, gl. 46.110 trgovina na debelo s kmetijskimi... ir
belo s pripravljenimi... pripravljenimi jedmi, gl. 46.380 trgovina na drobno s pripravljenimi... Skladiščenje nevarnih kemikalij Trgovina
pres, gl. 32.500 trgovina na debelo s farmacevtskimi... farmacevtskimi preparati, gl. 46.460 trgovina na drobno s farmacevtskimi... predel
na debelo z ostanki... in tehnologijo G TRGOVINA; VZDRŽEVANJE IN POPRAVILA... MOTORNIH VOZIL 45 Trgovina z motornimi vozili in... popravil
in na... Sem ne spada: trgovina na debelo ali drobno... avtomobilov na drobno 45.190 Trgovina z drugimi motornimi vozili... vozili Sem spa
trično 45.310 Trgovina na debelo z rezervnimi... vozila Sem spada: trgovina na debelo s plašči... za motorna vozila 45.320 Trgovina na dro
obno z motornimi... avtodelov preko interneta 45.400 Trgovina, vzdrževanje in popravila... popravila motornih koles; trgovina z njihovimi
... za motorna kolesa trgovina z motornimi kolesi in... Sem ne spada: trgovina na debelo s kolesi... opremo zanja, gl. 46.490 trgovina na
```

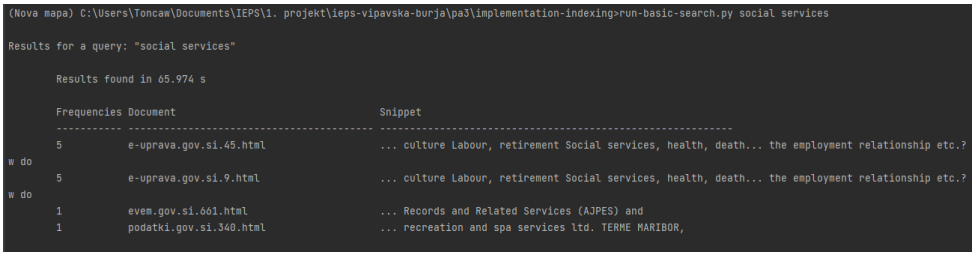
Slika 4: trgovina- basic search

3.0.3 Poizvedba 3: social services

Poizvedba vrne 4 rezultate.



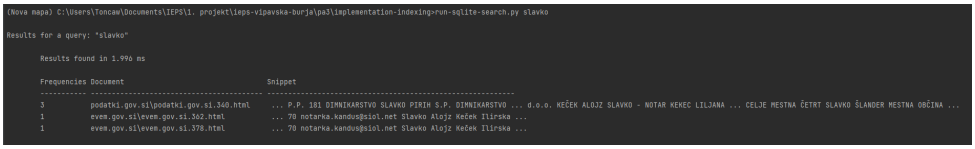
Slika 5: social services - sql search



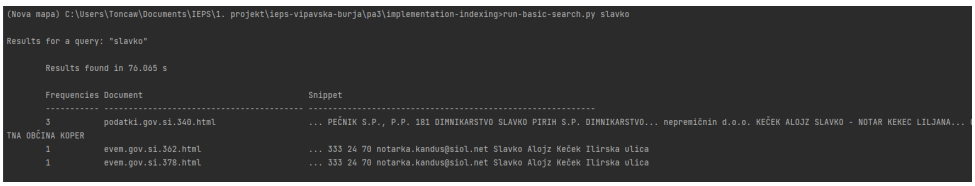
Slika 6: social services- basic search

3.0.4 Poizvedba 4: slavko

Poizvedba vrne 3 rezultate.



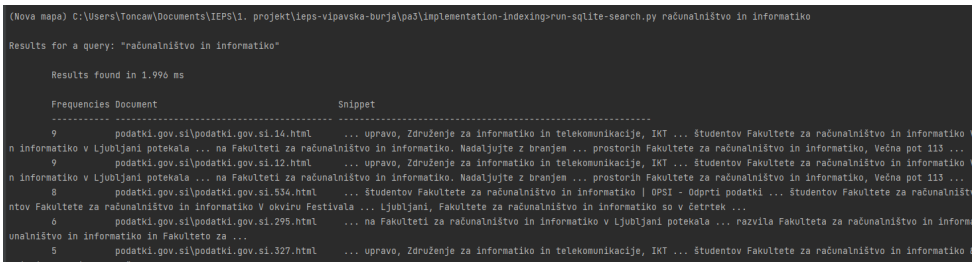
Slika 7: slavko - sql search



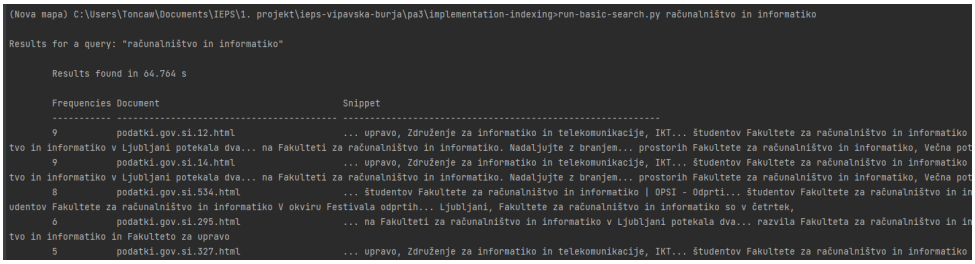
Slika 8: slavko - basic search

3.0.5 Poizvedba 5: računalništvo in informatiko

Poizvedba vrne 18 rezultatov. Na slikah je prikazanih 5 rezultatov z največjo frekvenco.



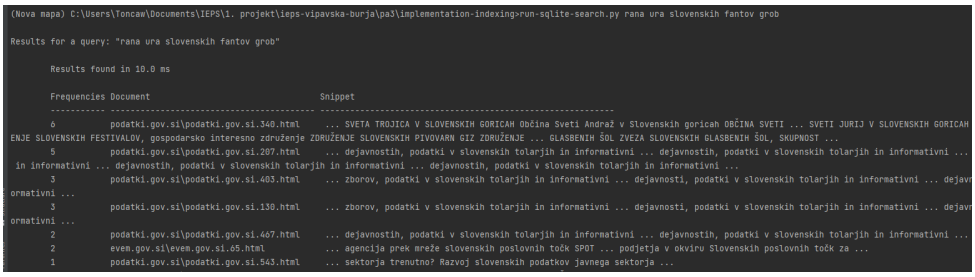
Slika 9: računalništvo in informatiko - sql search



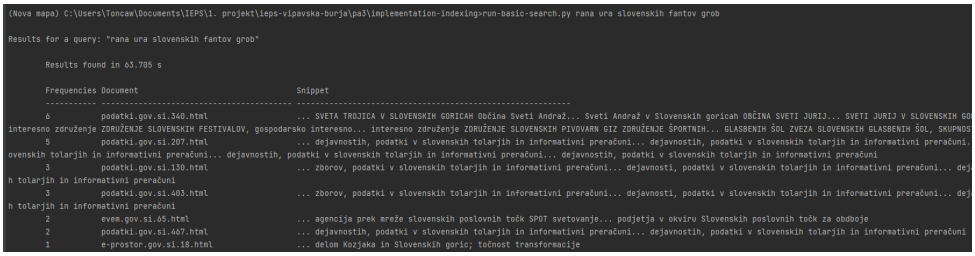
Slika 10: računalništvo in informatiko - basic search

3.0.6 Poizvedba 6: rana ura slovenskih fantov grob

Poizvedba vrne 29 rezultatov. Na slikah je prikazanih 7 rezultatov z največjo frekvenco.



Slika 11: rana ura slovenskih fantov grob - sql search



Slika 12: rana ura slovenskih fantov grob - basic search

Tu naj dodava, da je na priloženih slikah le odsek rezultatov poizvedbe in ponekod je mogoče, da vrstni red strani z enakim številom frekvence pojavitve ni enak pri iskanju z uporabo baze in ”basic search”