

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 besede 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 indeksom 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 poizvedbe in če hkrati ni v stop 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

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

Poizvedba vrne 2060 rezultatov.

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

```
Results for a query: "predelovadne dejavnosti"
```

```
Results found in 13.142 ms
```

Frequencies	Document	Snippet
1291	even.gov.si/even.gov.si.371.html	... iskanje ustrezne šifre dejavnosti /storitve in informacij o pogojih za opravljanje dejavnosti: Pridelava semenskega ... pogojih za opravljanje dejavnosti: Pridelava kmetijskih ... pogojih za opravljanje dejavnosti: Pridelava semenskega materiala ... pogojih za opravljanje dejavnosti: Pridelava in distribucija ... pogojih za opravljanje dejavnosti: Kih pridelekov ... 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

```
(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      even.gov.si.371.html                        ... iskanje ustrezne šifre dejavnosti /storitve in informacij o... pogojih za opravljanje
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: Pri
h pridelkov... pogojih za opravljanje dejavnosti: Pridelava semenskega materiala... aster) Pogoji: · Lista dejavnosti, ki se običajno... obrtni način
egistracija te dejavnosti poslovni subjekt 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
```

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 na d
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; VZORŽEVANJE 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 motorni
.490 trgovina na drobno s ... 46 Posredništvo in trgovina na debelo, razen ... Sem ne spada: trgovina na debelo za ... Sem ne spada: trgovina
```

```
(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                  ... zadrudnih organizacij, gl. 46.110 trgovina na debelo s kmetijskimi...
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... prede
na debelo z ostanki... in tehnologijo 6 TRGOVINA; VZDRŽEVANJE IN POPRAVILA... MOTORNIM VOZIL 45 Trgovina z motornimi vozili in... popravi
in na... Sem ne spada: trgovina na debelo ali drobno... avtomobilov na drobno 45.190 Trgovina z drugimi motornimi vozili... vozili Sem sp
trično 45.310 Trgovina na debelo z rezervnimi... vozila Sem spada: trgovina na debelo s plašči... za motorna vozila 45.320 Trgovina na dr
obno z motornimi... avto delov 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
```

2

3.0.3 Poizvedba 3: social services

Poizvedba vrne 4 rezultate.

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

Results for a query: "social services"

Results found in 3.997 ms

Frequencies Document Snippet
-----
5 e-uprava.gov.si\e-uprava.gov.si.9.html ... culture Labour, retirement Social services, health, death Taxes ... employment relationship etc.?
tance? How do ...
5 e-uprava.gov.si\e-uprava.gov.si.45.html ... culture Labour, retirement Social services, health, death Taxes ... employment relationship etc.?
tance? How do ...
1 podatki.gov.si\podatki.gov.si.340.html ... recreation and spa services ltd. TERME MARIBOR ...
1 even.gov.si\even.gov.si.601.html ... Records and Related Services (AJPES) and the ...
```

Slika 5: social services - sql search

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

Results for a query: "social services"

Results found in 65.974 s

Frequencies Document Snippet
-----
5 e-uprava.gov.si.45.html ... culture Labour, retirement Social services, health, death... the employment relationship etc.?
u do
5 e-uprava.gov.si.9.html ... culture Labour, retirement Social services, health, death... the employment relationship etc.?
u do
1 even.gov.si.601.html ... Records and Related Services (AJPES) and
1 podatki.gov.si.340.html ... recreation and spa services ltd. TERME MARIBOR,
```

Slika 6: social services- basic search

3.0.4 Poizvedba 4: slavko

Poizvedba vrne 3 rezultate.

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

Results for a query: "slavko"

Results found in 1.990 ms

Frequencies Document Snippet
-----
3 podatki.gov.si\podatki.gov.si.340.html ... P.P. 181 DIMNIKARSTVO SLAVKO PIRIH S.P. DIMNIKARSTVO ... d.o.o. KEČEK ALOJZ SLAVKO - NOTAR KEČEK LILJANA ... CELJE MESTNA ŽETRT SLAVKO ŠLANDER MESTNA OBČINA ...
1 even.gov.si\even.gov.si.362.html ... 70 notarka.kandus@siol.net Slavko Alojz Kečec Ilirska ...
1 even.gov.si\even.gov.si.378.html ... 70 notarka.kandus@siol.net Slavko Alojz Kečec Ilirska ...
```

Slika 7: slavko - sql search

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

Results for a query: "slavko"

Results found in 76.045 s

Frequencies Document Snippet
-----
3 podatki.gov.si.340.html ... PEČNIK S.P., P.P. 181 DIMNIKARSTVO SLAVKO PIRIH S.P. DIMNIKARSTVO... nepremičnin d.o.o. KEČEK ALOJZ SLAVKO - NOTAR KEČEK LILJANA...
TMA OBČINA KOPER
1 even.gov.si.362.html ... 333 24 70 notarka.kandus@siol.net Slavko Alojz Kečec Ilirska ulica
1 even.gov.si.378.html ... 333 24 70 notarka.kandus@siol.net Slavko Alojz Kečec Ilirska ulica
```

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.

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

Results for a query: "računalništvo in informatiko"

Results found in 1.996 ms

Frequencies	Document	Snippet
9	podatki.gov.si/podatki.gov.si.14.html	... upravo, Združenje za informatiko in telekomunikacije, IKT ... študentov Fakultete za računalništvo in informatiko
9	podatki.gov.si/podatki.gov.si.12.html	... upravo, Združenje za informatiko in telekomunikacije, IKT ... študentov Fakultete za računalništvo in informatiko
8	podatki.gov.si/podatki.gov.si.534.html	... študentov Fakultete za računalništvo in informatiko OPSI - Odprti podatki ... študentov Fakultete za računalništvo in informatiko
6	podatki.gov.si/podatki.gov.si.295.html	... na Fakulteti za računalništvo in informatiko v Ljubljani potekala ... razvila Fakulteta za računalništvo in informatiko
5	podatki.gov.si/podatki.gov.si.327.html	... upravo, Združenje za informatiko in telekomunikacije, IKT ... študentov Fakultete za računalništvo in informatiko

Slika 9: računalništvo in informatiko - sql search

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

Results for a query: "računalništvo in informatiko"

Results found in 04.764 s

Frequencies	Document	Snippet
9	podatki.gov.si.12.html	... upravo, Združenje za informatiko in telekomunikacije, IKT... študentov Fakultete za računalništvo in informatiko
9	podatki.gov.si.14.html	... upravo, Združenje za informatiko in telekomunikacije, IKT... študentov Fakultete za računalništvo in informatiko
8	podatki.gov.si.534.html	... študentov Fakultete za računalništvo in informatiko OPSI - Odprti... študentov Fakultete za računalništvo in informatiko
6	podatki.gov.si.295.html	... na Fakulteti za računalništvo in informatiko v Ljubljani potekala dva... razvila Fakulteta za računalništvo in informatiko
5	podatki.gov.si.327.html	... upravo, Združenje za informatiko in telekomunikacije, IKT... študentov Fakultete za računalništvo in informatiko

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.

```
(Nova mapa) C:\Users\Toncaw\Documents\IEPS\1. projekt\ieps-vipavska-burja\pa3\implementation-indexing-run-sqlite-search.py rana ura slovenskih fantov grob
```

Results for a query: "rana ura slovenskih fantov grob"

Results found in 10.0 ms

Frequencies	Document	Snippet
6	podatki.gov.si/podatki.gov.si.340.html	... SVETA TROJICA V SLOVENSKIH GORICAH Občina Sveti Andraž v Slovenskih goricah OBČINA SVETI ... SVETI JURIJ V SLOVENSKIH GORICAH
5	podatki.gov.si/podatki.gov.si.207.html	... dejavnosti, podatki v slovenskih tolarjih in informativni ... dejavnosti, podatki v slovenskih tolarjih in informativni ...
3	podatki.gov.si/podatki.gov.si.403.html	... Zborov, podatki v slovenskih tolarjih in informativni ... dejavnosti, podatki v slovenskih tolarjih in informativni ... dejavnosti
3	podatki.gov.si/podatki.gov.si.130.html	... Zborov, podatki v slovenskih tolarjih in informativni ... dejavnosti, podatki v slovenskih tolarjih in informativni ... dejavnosti
2	podatki.gov.si/podatki.gov.si.407.html	... dejavnosti, podatki v slovenskih tolarjih in informativni ... dejavnosti, podatki v slovenskih tolarjih in informativni ...
2	even.gov.si/even.gov.si.05.html	... agencija prek mreže slovenskih poslovnih točk SPOT ... podjetja v okviru Slovenskih poslovnih točk za ...
1	podatki.gov.si/podatki.gov.si.543.html	... sektorja trenutno? Razvoj slovenskih podatkov javnega sektorja ...

Slika 11: rana ura slovenskih fantov grob - sql search

```
(Nova mapa) C:\Users\Toncaw\Documents\IEPS\1. projekt\ieps-vipavska-burja\pa3\implementation-indexing-run-basic-search.py rana ura slovenskih fantov grob
```

Results for a query: "rana ura slovenskih fantov grob"

Results found in 63.705 s

Frequencies	Document	Snippet
6	podatki.gov.si.340.html	... SVETA TROJICA V SLOVENSKIH GORICAH Občina Sveti Andraž... Sveti Andraž v Slovenskih goricah OBČINA SVETI JURIJ... SVETI JURIJ V SLOVENSKIH GORICAH
5	podatki.gov.si.207.html	... dejavnosti, podatki v slovenskih tolarjih in informativni preračuni... dejavnosti, podatki v slovenskih tolarjih in informativni preračuni
3	podatki.gov.si.130.html	... Zborov, podatki v slovenskih tolarjih in informativni preračuni... dejavnosti, podatki v slovenskih tolarjih in informativni preračuni
3	podatki.gov.si.403.html	... Zborov, podatki v slovenskih tolarjih in informativni preračuni... dejavnosti, podatki v slovenskih tolarjih in informativni preračuni
2	even.gov.si.05.html	... agencija prek mreže slovenskih poslovnih točk SPOT svetovanje... podjetja v okviru Slovenskih poslovnih točk za območje
2	podatki.gov.si.407.html	... dejavnosti, podatki v slovenskih tolarjih in informativni preračuni... dejavnosti, podatki v slovenskih tolarjih in informativni preračuni
1	e-prostor.gov.si.18.html	... delom Korjaka in Slovenskih gorici; točnost transformacije

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"