Programming Assignment 3

Implementing a data index and more

Jaka Kokošar, Danijel Maraž, and Toni Kocjan

Fakulteta za Računalništvo in Informatiko UL dm9929@student.uni-lj.si, jk0902@student.uni-lj.si, tk3152@student.uni-lj.si

Povzetek The article covers the work done in the scope of the third programming assignment as part of the subject web information extraction and retrieval.

Keywords: Data Processing Indexing Retrieval

1 Uvod

Po tem, ko smo zbrali spletne strani v prvi nalogi in jih temeljito razčlenili na podatke, ki nas zanimajo v drugi, smo bili pripravljeni nadaljevati z zadnjim korakom, ki je izdelava indeksa podatkov in izvajanje poizvedb.

2 Data Processing

Glavna funkcija *preprocess* prejme kot argument rezultat funkcije *text*, ki sama prejme našo surovo html vsebino in tej odstrani nepotrebne tehnične html oznake. Preprocess nato:

- S funkcijo *remove_punctuation* odstrani ločila in več
- Z *nltk.tokenize.word_tokenize* pretvori v vrsto besednih značk
- Odstrani se značke, ki niso alfabetične
- Vse velike začetnice se pretvorijo v male
- Odstrani se značke, ki so *stopword*
- S pomočjo pretvorbe v podatkovno strukturo množice se odstranijo duplikati

Nato ta vrne seznam ostalih značk.

3 Indexing

Funkcija *initiating_indexing* se požene in začne meriti čas gradnje indeksa. Na koncu izpiše na standardni izhod celoten porabljen čas. Glavno nalogo indeksiranja opravlja razred *BetterThanGoogle* nad katerim se pokliče funkcijo *create_index* in se mu kot argument poda relativno pot do datotek za sestavo indeksa.

3.1 BetterThanGoogle

3.2 Create_index

Funkcija kot argumenta prejme instanci razredov *Preprocessor* (ta služi za procesiranje besedila po opisu iz poglavja Data Processing) in *DBHandler* (ta služi za interakcijo z bazo). Nato pridobimo ime datoteke ter vsebino s pomočjo naše abstrakcije korpusov (*file_name* in *document*). Za tem iteriramo skozi vsak par ter spremenljivko *document* ustrezno obdelamo z razredom *preprocessor* (glej *__call__* od preprocessor). Za vsako značko, ki jo vrne preprocessor:

- Najdemo vse njene pojavitve v besedilu (*find_occurrences*)
- Pod pogojem, da smo našli vsaj eno pojavitev se izvede faza vnosa v indeks
- V indeks vnesemo ime značke, ime datoteke, število pojavitev značke, ter niz posameznih pojavitev ločen z vejico

Vredno je tudi omeniti, da program v log ves čas izpisuje koliko datotek je obdelal do sedaj in koliko mu jih še manjka.

3.3 Statistika indeksa

10 najpogostejših besed

- ro data/evem.gov.si/evem.gov.si.371.html 6968
- el data/evem.gov.si/evem.gov.si.371.html 4466
- st data/podatki.gov.si/podatki.gov.si.340.html 3582
- vod data/evem.gov.si/evem.gov.si.371.html 3111
- go data/evem.gov.si/evem.gov.si.371.html 3048
- rs data/evem.gov.si/evem.gov.si.371.html 2919
- tv data/evem.gov.si/evem.gov.si.371.html 2862
- ir data/evem.gov.si/evem.gov.si.371.html 2442
- sp data/evem.gov.si/evem.gov.si.371.html 2424
- proizvodnja data/evem.gov.si/evem.gov.si.371.html 2268

10 najredkejših besed

- soglašate data/e-prostor.gov.si/e-prostor.gov.si.192.html 1
- ministrstva data/e-prostor.gov.si/e-prostor.gov.si.192.html 1
- vlade data/e-prostor.gov.si/e-prostor.gov.si.192.html 1
- posredovanju data/e-prostor.gov.si/e-prostor.gov.si.192.html 1
- uredbo data/e-prostor.gov.si/e-prostor.gov.si.192.html 1
- ponovni data/e-prostor.gov.si/e-prostor.gov.si.192.html 1
- zaračunavanja data/e-prostor.gov.si/e-prostor.gov.si.192.html 1
- rtcm data/e-prostor.gov.si/e-prostor.gov.si.192.html 1
- klicem data/e-prostor.gov.si/e-prostor.gov.si.192.html 1
- upravnih data/e-prostor.gov.si/e-prostor.gov.si.192.html 1

Celotno število indeksiranih besed: 32248

4 Data Retrieval

Funkcija *initiating_search* prejme niz za katerega želimo iskati pojavitve v trenutnem indeksu. Ta ustvari nov objekt *SearchEngine* in mu poda instanco *DBHandler*. Potem se za dejansko iskanje na ustvarjenem objektu kliče funkcijo *perform_query*. Nato program izpiše na zaslon rezultate in čas porabljen za poizvedbo.

4.1 SearchEngine

Razred *SearchEngine* kot argument prejme *db_handler* (preko katerega se izvajajo vse interakcije z bazo).

Perform_query Funkcija kot argument prejme niz besed ločenih s presledki. Nato iz teh ustvari seznam besed, ter vsaki besedi velike črke zamenja z malimi. Za vsako se nato naredi poizvedba v bazi iz tabele *Posting* in s funkcijo *_find_occurrences_in_file* in več zankami ustvari terko *QueryResults(beseda, snippeti pojavitev besede)*. Na koncu funkcija vrne drugo terko s številom pojavitev na prvem mestu in seznamom terk *QueryResults* na drugem.

Hitrost Vredno je omeniti, da ob samem iskanju za pridobitev snippetov besedila vsakič znova obdelamo datoteke. Posledično so nekatere poizvedbe zelo počasne. Dejstva se zavedamo in menimo, da bi se dalo ravno tukaj bistveno nadgraditi sistem s predhodnim pomnjenjem snippetov.

5 Rezultati poizvedb

5.1 predelovalne dejavnosti

```
Searching time: 22.426688194274902
Found 6367 results for "predelovalne dejavnosti"

Found 1570 results in
   "data/evem.gov.si/evem.gov.si.371.html":
0: '... – \n \n * **C** PREDELOVALNE DEJAVNOSTI\n\n
   ...'
1: '... ruge raznovrstne predelovalne dejavnosti\n\n ...'
2: '... gje ščnerazvrene predelovalne dejavnosti\n\n ...'
3: '... v industrijskem predelovalnem procesu, gl ...'
4: '... e iz čpodroja C (Predelovalne dejavnosti)_ ...'
5: '... e ustrezne šifre dejavnosti /storitve in ...'
6: '... h za opravljanje dejavnosti.\n\nV iskalnik ...'
7: '... ih je 645 od 645 dejavnosti\n\nIzpisanih j ...'
8: '... zpisanih je od dejavnosti\n\n * **A** K ...'
9: '... h za opravljanje dejavnosti:\n\·n Pridelav ...'
10: '... h za opravljanje dejavnosti:\n\·n Pridelav ...'
11: '... h za opravljanje dejavnosti:\n\nPridelava ...'
12: '... h za opravljanje dejavnosti:\n\n ...'
13: '... h za opravljanje dejavnosti:\n\n ...'
14: '... h za opravljanje dejavnosti:\n\n ...'
15: '... h za opravljanje dejavnosti:\n\n ...'
16: '... h za opravljanje dejavnosti:\n\n ...'
17: '... **\n\n### · Lista dejavnosti, ki se čobia ...'
18: '... obrtni čnain\n\nZa dejavnosti, ki so ščuvr ...'
19: '... ščvrene na Listo dejavnosti, ki se čobia ...'
20: '... registraciji te dejavnosti poslovni sub ...'
```

5.2 trgovina

```
Searching time: 5.710252046585083
Found 1158 results for "trgovina"

Found 368 results in
   "data/evem.gov.si/evem.gov.si.371.html":
```

```
0: '... .110_\n        * _trgovina na debelo s  ...'
1: '... .890_\n        * _trgovina na debelo z  ...'
2: '... .890_\n        * _trgovina na debelo s  ...'
3: '... .380_\n        * _trgovina na drobno s  ...'
4: '... alij\n          * Trgovina na debelo z  ...'
5: '... jami\n          * Trgovina na drobno z  ...'
6: '... .500_\n        * _trgovina na debelo s  ...'
7: '... .460_\n        * _trgovina na drobno s  ...'
8: '... a čščienje tal v trgovinah in industri ...'
9: '... .320_\n        * _trgovina (odkup in pr ...'
10: '... .220_\n        * _trgovina na debelo z  ...'
11: '... o  \n  \n  * **G** TRGOVINA; ŽVZDREVANJE ...'
12: '... IL\n\n    * **45** Trgovina z motornimi  ...'
13: '...      * **45.110** Trgovina z avtomobili ...'
14: '... da:**\n\n        * trgovina na debelo in ...'
15: '... :_**\n\n        * _trgovina na debelo al ...'
16: '...      * **45.190** Trgovina z drugimi mo ...'
17: '... da:**\n\n        * trgovina na debelo al ...'
18: '... :_**\n\n        * _trgovina na debelo al ...'
19: '...      * **45.310** Trgovina na debelo z  ...'
20: '... da:**\n\n        * trgovina na debelo s  ...'
```

5.3 social services

```
Searching time:  1.4749388694763184
Found 12 results for "social services"

Found 5 results in
   "data/e-uprava.gov.si/e-uprava.gov.si.45.html":
0: '... retirement\n  * Social services, heal ...'
1: '... hip etc.?\n\n##  Social services, heal ...'
2: '... tain financial social assistance? Ho ...'
3: '... ent\n  * Social services, health, death ...'
4: '... .?\n\n##  Social services, health, death ...'

Found 5 results in
   "data/e-uprava.gov.si/e-uprava.gov.si.9.html":
0: '... retirement\n  * Social services, heal ...'
1: '... hip etc.?\n\n##  Social services, heal ...'
```

```
2: '... tain financial social assistance? Ho ...'
3: '... ent\n  * Social services, health, death ...'
4: '... .?\n\n##  Social services, health, death ...'

Found 1 results in
   "data/evem.gov.si/evem.gov.si.661.html":
0: '... ords and Related Services (AJPES) and  ...'

Found 1 results in
   "data/podatki.gov.si/podatki.gov.si.340.html":
0: '... creation and spa services ltd.\n\nTERME  ...'
```

5.4 MJU

```
Searching time:  0.6603918075561523
Found 28 results for "MJU"

Found 5 results in
   "data/podatki.gov.si/podatki.gov.si.295.html":
0: '... 017/18** , ki ga MJU organizira s ...'
1: '... \n\n![](http://www.mju.gov.si/filea ...'
2: '... gov.si/fileadmin/mju.gov.si/pageu ...'
3: '... )\n![](http://www.mju.gov.si/filea ...'
4: '... gov.si/fileadmin/mju.gov.si/pageu ...'

Found 3 results in
   "data/podatki.gov.si/podatki.gov.si.351.html":
0: '... ronska špota: gp.mju@gov.si, \n     ...'
1: '... a: ispap-podatki.mju@gov.si\n    * ...'
2: '... l)\n\nhtml\n\n###### MJU\n\n**Podrobnos ...'

Found 2 results in
   "data/evem.gov.si/evem.gov.si.68.html":
0: '... za javno upravo (MJU), ki na podl ...'
1: '... racija posreduje\nMJU, ki uredi za ...'

Found 2 results in
   "data/podatki.gov.si/podatki.gov.si.105.html":
0: '... ronska špota: gp.mju@gov.si\n    * ...'
```

```
1: '... 1)\n\nhtml\n\n###### MJU\n\n**Podrobnos ...'
```

5.5 državni oblak

```
Searching time: 42.03738808631897
Found 3873 results for "ždravni oblak"

Found 29 results in
   "data/podatki.gov.si/podatki.gov.si.106.html":
0: '... \n\nOrganizacija:\n\ŽnDRAVNI ZBOR REPUBLI ...'
1: '... IKE SLOVENIJE\n\n# ŽDRAVNI ZBOR REPUBLI ...'
2: '... jubljene zbirke\n\žnDravni organi\n\nJavn ...'
3: '... de __78 ogledov\n\ŽnDRAVNI ZBOR REPUBLI ...'
4: '... ranjem ...\n\nXML\n\žnDravni organi\n\nJavn ...'
5: '... ka __27 ogledov\n\ŽnDRAVNI ZBOR REPUBLI ...'
6: '... ranjem ...\n\nXML\n\žnDravni organi\n\nJavn ...'
7: '... ra __19 ogledov\n\ŽnDRAVNI ZBOR REPUBLI ...'
8: '... ranjem\n...\n\nXML\n\žnDravni organi\n\nJavn ...'
9: '... ov __33 ogledov\n\ŽnDRAVNI ZBOR REPUBLI ...'
10: '... edila, ki jih je žDravni zbor sprejel ...'
11: '... ranjem ...\n\nXML\n\žnDravni organi\n\nJavn ...'
12: '... ra __37 ogledov\n\ŽnDRAVNI ZBOR REPUBLI ...'
13: '... ranjem ...\n\nXML\n\žnDravni organi\n\nJavn ...'
14: '... ka __22 ogledov\n\ŽnDRAVNI ZBOR REPUBLI ...'
15: '... ranjem ...\n\nXML\n\žnDravni organi\n\nJavn ...'
16: '... ti __17 ogledov\n\ŽnDRAVNI ZBOR REPUBLI ...'
17: '... aktov, ki jih je\žnDravni zbor sprejel ...'
18: '... ranjem ...\n\nXML\n\žnDravni organi\n\nJavn ...'
19: '... ra __16 ogledov\n\ŽnDRAVNI ZBOR REPUBLI ...'
20: '... ranjem ...\n\nXML\n\žnDravni organi\n\nJavn ...'
```

5.6 lahko tudi komisija

```
Searching time: 38.3415310382843
Found 2205 results for "lahko tudi komisija"

Found 287 results in
   "data/evem.gov.si/evem.gov.si.371.html":
0: '... st na kmetiji se lahko čzane izvaja ...'
1: '... st na kmetiji se lahko čzane izvaja ...'
```

```
2: '... bliki Sloveniji, lahko lovijo kot l ...'
3: '... o.\n\nLovski čuvaj lahko postane poln ...'
4: '... kolarjenjem** se lahko ukvarja oseb ...'
5: '... st na kmetiji se lahko čzane izvaja ...'
6: '... st na kmetiji se lahko čzane izvaja ...'
7: '... ev in školjk, ki lahko poteka v obr ...'
8: '... šiki gospodar je lahko vsak polnole ...'
9: '... ektroribolova je lahko vsak polnole ...'
10: '... šRibiki čuvaj je lahko vsak polnole ...'
11: '... škega šokolia je lahko vsak polnole ...'
12: '... st na kmetiji se lahko čzane izvaja ...'
13: '... šč izkorianje se lahko podeli tudi ...'
14: '... darskih del** je lahko pravna ali f ...'
15: '... rudarskih del je lahko posameznik, ...'
16: '... šč izkorianje se lahko podeli tudi ...'
17: '... darskih del** je lahko pravna ali f ...'
18: '... rudarskih del je lahko posameznik, ...'
19: '... šč izkorianje se lahko podeli tudi ...'
20: '... darskih del** je lahko pravna ali f ...'
```

5.7 slovenija

```
Found 67 results in
   "data/podatki.gov.si/podatki.gov.si.340.html":
0: '... jo d.o.o.\n\nCIPRA–SLOVENIJA Zavod za var ...'
1: '... LNI KOMITE PIARC SLOVENIJA, giz; v angl ...'
2: '... ŠKOCJANSKE JAME, Slovenija\n\nPartim, fin ...'
3: '... \nRADIOTELEVIZIJA SLOVENIJA javni zavod, ...'
4: '... aciji\n\nREPUBLIKA SLOVENIJA\n\nREPUBLIKA S ...'
5: '... ENIJA\n\nREPUBLIKA SLOVENIJA, MINISTRSTVO ...'
6: '... NOSTI\n\nREPUBLIKA SLOVENIJA, MINISTRSTVO ...'
7: '... ŽELJA\n\nREPUBLIKA SLOVENIJA UPRAVNA ENOT ...'
8: '... ŠČINA\n\nREPUBLIKA SLOVENIJA UPRAVNA ENOT ...'
9: '... ŽEICE\n\nREPUBLIKA SLOVENIJA UPRAVNA ENOT ...'
10: '... CELJE\n\nREPUBLIKA SLOVENIJA UPRAVNA ENOT ...'
11: '... KNICA\n\nREPUBLIKA SLOVENIJA UPRAVNA ENOT ...'
12: '... OMELJ\n\nREPUBLIKA SLOVENIJA UPRAVNA ENOT ...'
13: '... ŽMALE\n\nREPUBLIKA SLOVENIJA UPRAVNA ENOT ...'
14: '... OGRAD\n\nREPUBLIKA SLOVENIJA UPRAVNA ENOT ...'
```

```
15: '... DGONA\n\nREPUBLIKA SLOVENIJA UPRAVNA ENOT ...'
16: '... UPLJE\n\nREPUBLIKA SLOVENIJA UPRAVNA ENOT ...'
17: '... STNIK\n\nREPUBLIKA SLOVENIJA UPRAVNA ENOT ...'
18: '... DRIJA\n\nREPUBLIKA SLOVENIJA UPRAVNA ENOT ...'
19: '... TRICA\n\nREPUBLIKA SLOVENIJA UPRAVNA ENOT ...'
20: '... IZOLA\n\nREPUBLIKA SLOVENIJA UPRAVNA ENOT ...'
```

5.8 davki

```
Searching time: 41.975666999816895
Found 2767 results for "davki"

Found 39 results in
   "data/evem.gov.si/evem.gov.si.32.html":
0: '... ega čobrauna na eDavkihOstalo\n\n ...'
1: '... ega\čnobrauna na čeDavkihFizine oseb ...'
2: '... ega čobrauna na eDavkihPravne osebe ...'
3: '... ega čobrauna na eDavkihOstalo\n\n ...'
4: '... ega\čnobrauna na čeDavkihFizine oseb ...'
5: '... ega čobrauna na eDavkihPravne osebe ...'
6: '... ega\čnobrauna na čeDavkihFizine oseb ...'
7: '... ega čobrauna na eDavkihPravne osebe ...'
8: '... ega čobrauna na eDavkihOstalo\n\n ...'
9: '... ega\čnobrauna na čeDavkihFizine oseb ...'
10: '... ega čobrauna na eDavkihOstalo\n\n ...'
11: '... ega\čnobrauna na čeDavkihFizine oseb ...'
12: '... ega čobrauna na eDavkihPravne osebe ...'
13: '... ega čobrauna na eDavkihOstalo\n\n ...'
14: '... ega čobrauna na eDavkihPravne osebe ...'
15: '... ega čobrauna na eDavkihOstalo\n\n ...'
16: '... ega\čnobrauna na čeDavkihFizine oseb ...'
17: '... ega\čnobrauna na čeDavkihFizine oseb ...'
18: '... ega\čnobrauna na eDavkihPravne osebe ...'
19: '... ega čobrauna na eDavkihPravne osebe ...'
20: '... ega čobrauna na čeDavkihFizine oseb ...'
```