Poročilo 3. Seminarske naloge

Avtorji: Nal Lukšič, Aljaž Rupar, Gregor Ažbe

Mentor: assist. prof. dr. Slavko Žitnik

Uvod

Seminarska naloga je sestavljena iz dveh delov. V prvem delu naloge smo implementirali program za pridobivanje in indeksiranje besed iz dokumentov, v drugem pa 2 rešitvi za iskanje besed z pomočjo indeksov. Rešitev se uporablja za učinkovito iskanje besed ali fraz po dokumentih.

Tehnologije

Rešitev je izdelana v Python 3.8 v rozvojnem okolju Pycharm. Za implementacijo skript smo poleg že integriranih paketov uporabili še module Beauftiful Soup 4, sqlite3, nltk, re in defaultdict.

Specifikacije rešitve

1. Indeksiranje dokumentov

Prvi del seminarske naloge smo rešili z programom, ki sprejme različne html datoteke, jih očisti in besede shrani v SQLite podatkovno bazo po shemi in v formatu danem na spletni učilnici.

Program je narejen na način, da najde vse mape v »input-index« in iterira čez vse vsebovane HTML dokumente.Iz vsakega dokumenta najprej s pomočjo knjižnice Beautiful Soup odstranimo komentarje, css, skripte, linke in iframe. Čez očiščen dokument nato zaženemo funkcijo process_text, ki tokenizira text z uporabo nltk. word_tokenize in zmanjša velike začetnice. Dobljenemu besedilu nato še odstranimo stop besede podane na spletni učilnici – stop-words-slovene. Končne tokene nato shranimo v defaultdict za lažjo pripravo na poznejši zapis v podatkovno bazo.

Dobljene tokene nato shranimo v podatkovno bazo v tabeli IndexWord in Posting po principu opisanem na spletni učilnici.

2. Iskanje po dokumentih

Iskanje fraze v podatkovni bazi smo implementirali na 2 načina in sicer z že zgrajenim indeksom in z sprotno iskanje:

- Z že zgrajenim indeksom (run-sqlite-search.py)

Skripta deluje tako, da mu kot argumente podamo besede, program pa nato izpiše rezultate v obliki kot je opisano na spletni strani.

Po zagonu program se najprej vzpostavi povezava z podatkovno bazo. Nato tokeniziramo, zmanjšamo velike začetnice in odstranimo stop besede iz podanih argumentov. Naslednji korak je iskanje očiščenega vhoda v podatkovni bazi z sql poizvedbo. Najdene rešitve nato najdemo v dokumentih v katerih se nahajajo. Te dokumente tako kot pri indeksiranju dokumentov očistimo in najdemo okolico rešitve, katero izpišemo na standardni izhod.

- Brez zgrajenega indeksa (run-basic-search.py)

Ta del je združena kombinacija prvih dveh programov, ker najprej očisti vhod, nato iterira čez dokumente in uredi njihovo vsebino, kot je opisano v indeksiranje dokumentov, nato pa vrne rezultate na standardni izhod.

Rezultati

Rezultati za poizvedbe:

Poizvedba	Rezultat
predelovalne dejavnosti	Results for a query: predelovalne dejavnosti Results found in 0.006409500000000179s. --- FrequencyDocuments Snippet --- 75 evem.gov.si.377.html Defektolog v zdravstveni dejavnosti ... Dietetik v zdravstveni dejavnosti ... v zd 40 evem.gov.si.452.html Druge storitvene dejavnosti, drugje nerazvrščene ... Dejavnosti ... Druge storitvene 40 podatki.gov.si.340.html NOSILEC DOPOLNILNE dejavnosti NA KMETIJI ... CENTER INTERESNIH dejavnosti PTUJ ... ritve, 31 evem.gov.si.653.html za opravljanje dejavnosti specializirane prodajalne ... ali televizijske dejavnosti 30 evem.gov.si.398.html na opravljanje dejavnosti (npr.: pripravljalna ... z opravljanjem dejavnosti v Slove 29 evem.gov.si.72.html dohodka iz dejavnosti ... Davek od dohodka iz dejavnosti ... Davek od dohodka iz dej je 25 evem.gov.si.442.html dejavnosti za nego ... Dejavnosti ... Dejavnosti za nego telesa (96.040) ... Dejavno 20 evem.gov.si.460.html Drugje nerazvrščene predelovalne dejavnosti (32.990) ... nerazvrščene predelovalne d 18 evem.gov.si.265.html (10.110) / dejavnosti / eVEM ... Dejavnosti ... zajema dejavnosti in storitve, ... 18 evem.gov.si.28.html opravljanje gospodarske dejavnosti. ... za posamezne dejavnosti ali posamezne ... in 17 evem.gov.si.266.html (10.130) / dejavnosti / eVEM ... Dejavnosti ... zajema dejavnosti in storitve, ...

trgovina	Results for a query: trgovina Results found in 0.008850400000000036s. --- FrequencyDocuments Snippet --- 96 evem.gov.si.651.html Druga trgovina na drobno ... Druga trgovina na drobno ... Druga t 92 evem.gov.si.21.html Trgovina ... Druga trgovina na drobno ... Druga trgovina na drobno . 82 podatki.gov.si.340.html A DENT, trgovina in storitve, ... ADRIA INVESTICIJE trgovina, posred 14 evem.gov.si.623.html trgovina na debelo ... Trgovina na debelo z drugimi izdelki široke p 13 evem.gov.si.329.html trgovina na debelo ... trgovina na debelo ... trgovina na debelo ... 13 evem.gov.si.630.html trgovina na drobno ... trgovina na drobno ... trgovina na drobno ... 11 evem.gov.si.320.html trgovina na debelo ... trgovina na debelo ... trgovina na debelo ... 11 evem.gov.si.327.html trgovina na debelo ... Trgovina na debelo z drugimi napravami in opr 11 evem.gov.si.622.html trgovina na debelo ... trgovina na debelo ... trgovina na debelo ...
social services	Results for a query: social services Results found in 0.008590499999999501s. --- FrequencyDocuments Snippet --- 5 e-uprava.gov.si.9.html Social services, health, death ... Social services, h 1 evem.gov.si.661.html and Related services (AJPES) and 1 podatki.gov.si.340.html and spa services ltd.
opis poročila	Results for a query: opis poročila Results found in 0.008617100000000377s. --- FrequencyDocuments Snippet --- 14 podatki.gov.si.379.html opis datoteke Organi_niso_porocali_na_dan_1._julij_2016_izpi 9 e-prostor.gov.si.30.html Opis ... Metapodatkovni opis ... Metapodatkovni opis ... M 7 e-prostor.gov.si.1.html Opis ... Poročila o trgu nepremičnin ... objavljati periodič 6 e-prostor.gov.si.11.html Opis ... Metapodatkovni opis ... Opis ... Opis »shape« for 5 e-prostor.gov.si.13.html opis meje, ... opis meje (razsodba arbitražnega sodišča), .. 5 e-prostor.gov.si.170.html Podatek in povezava na opis ... opis strukture podatkov ... 5 e-prostor.gov.si.7.html Podatek in povezava na opis ... opis strukture podatkov ... 4 e-prostor.gov.si.12.html Opis ... Opis elementov diagrama aktivnosti ... Metapodatkov 4 e-prostor.gov.si.150.html Opis projekta ... Kratek opis vsebine projekta je na voljo 4 e-prostor.gov.si.18.html Oznaka in kratek opis ... Oznaka in kratek opis ... Oznaka i 4 podatki.gov.si.66.html opis datoteke Porocilo+uspesnost+tehnicnih+pregledov+2015 ..
trg	

<table>
<tr><td></td><td>

```
        Results for a query: trg

        Results found in 0.00879120000000011s.

--------------------------------------------------
FrequencyDocuments                    Snippet
--------------------------------------------------
16      e-uprava.gov.si.36.html       2 TRBOVLJE Mestni trg 4  ... 2 TRBOVLJE Mestni trg 4  ... NOVA G
14      evem.gov.si.362.html          Trg zbora odposlancev 66 ... Glavni trg 41 ... Trg republike 3 .
14      evem.gov.si.378.html          Trg zbora odposlancev 66 ... Glavni trg 41 ... Trg republike 3 .
6       evem.gov.si.279.html          Priprava za trg in trženje ... Priprava za trg in trženje ... Pr
6       podatki.gov.si.340.html       AGENCIJA ZA trg VREDNOSTNIH PAPIRJEV ... SKUPNOST NOVI trg   ...
4       evem.gov.si.575.html          Priprava za trg in trženje ... Priprava za trg in trženje semens
3       evem.gov.si.264.html          pripravo za trg, uvozom oziroma ... Priprava za trg in trženje s
3       evem.gov.si.398.html          dostopa na trg dela? ... dostopa na trg dela, se
3       evem.gov.si.651.html           Dajanje tretiranih izdelkov na trg ... Priprava za trg in tržen
2       e-prostor.gov.si.2.html       slovenski nepremičninski trg leta... ... slovenski nepremičninsk
2       e-prostor.gov.si.33.html      slovenski nepremičninski trg leta... ... slovenski nepremičninsk
```

</td></tr>
<tr><td>aktualno</td><td>

```
        Results for a query: aktualno

        Results found in 0.008633699999999855s.

--------------------------------------------------
FrequencyDocuments                  Snippet
--------------------------------------------------
5       e-prostor.gov.si.162.html   E-prostor - Aktualno ... Aktualno ... Aktualno ... Aktualno ... Arhiv aktualno
5       e-prostor.gov.si.164.html   E-prostor - Aktualno ... Aktualno ... Aktualno ... Aktualno ... Arhiv aktualno
5       e-prostor.gov.si.189.html   E-prostor - Aktualno ... Aktualno ... Aktualno ... Aktualno ... Arhiv aktualno
5       e-prostor.gov.si.190.html   E-prostor - Aktualno ... Aktualno ... Aktualno ... Aktualno ... Arhiv aktualno
5       e-prostor.gov.si.191.html   E-prostor - Aktualno ... Aktualno ... Aktualno ... Aktualno ... Arhiv aktualno
5       e-prostor.gov.si.50.html    E-prostor - Aktualno ... Aktualno ... Aktualno ... Aktualno ... Arhiv aktualno
5       e-prostor.gov.si.98.html    E-prostor - Aktualno ... Aktualno ... Aktualno ... Aktualno ... Arhiv aktualno
5       e-prostor.gov.si.99.html    E-prostor - Aktualno ... Aktualno ... Aktualno ... Aktualno ... Arhiv aktualno
4       e-prostor.gov.si.101.html   E-prostor - Aktualno - arhiv ... Aktualno - arhiv ... Aktualno - arhiv ... Aktualno - arhiv
4       e-prostor.gov.si.102.html   E-prostor - Aktualno - arhiv ... Aktualno - arhiv ... Aktualno - arhiv ... Aktualno - arhiv
4       e-prostor.gov.si.103.html   E-prostor - Aktualno - arhiv ... Aktualno - arhiv ... Aktualno - arhiv ... Aktualno - arhiv
4       e-prostor.gov.si.176.html   E-prostor - Aktualno - arhiv ... Aktualno - arhiv ... Aktualno - arhiv ... Aktualno - arhiv
4       e-prostor.gov.si.177.html   E-prostor - Aktualno - arhiv ... Aktualno - arhiv ... Aktualno - arhiv ... Aktualno - arhiv
4       e-prostor.gov.si.178.html   E-prostor - Aktualno - arhiv ... Aktualno - arhiv ... Aktualno - arhiv ... Aktualno - arhiv
```

</td></tr>
</table>

Primerjava hitrosti:

Iskalni izraz	Z uporabo indeksov	Brez uporabe indeksov
trgovina	0.00885s	100.67s
Social services	0.00859s	95.24s
Predelovalne dejavnosti	0.00864s	93.53s
opis poročila	0.00861s	93.19s
trg	0.00879s	91.14s
aktualno	0.00863s	97.30s