Seminarska naloga 3 - Ekstrakcija podatkov iz spleta

Gal Bumbar, Jan Lampič, Maja Umek

Maj 2019

1 Uvod

Internet je za večino ljudi postal prvi vir, na katerega se obrnejo, kadar iščejo informacije o določenem izdelku ali storitvi. Spletni iskalniki so tako postali orodje, brez katerega danes praktično ne gre. Spletni iskalnik je programski sistem, ki je zasnovan za izvajanje spletnega iskanja, kar pomeni sistematično iskanje po svetovnem spletu za informacije, določene v poizvedbi za iskanje. V splošnem je sestavljen iz treh delov:

- Spletni pajek, ki išče in zapiše primerne spletne strani v svojo bazo.
- Indekser ali kazalnik, ki procesira vsebino spletne strani ter indeksira vsako besedo na posamezni spletni strani in jo shrani v poseben register, kjer so povezane besede in spletne strani.
- Iskalni procesor ki primerja vaše iskalne nize z registrom shranjenih besed in predlaga najbolj ustrezno spletno stran.

V prvih seminarski nalogi smo ustvarili spletnega pajka za prenos spletnih vsebin. V drugi smo si pogledali, kako iz dobljenih spletnih strani lahko izluščimo želene informacije, v tem delu pa si bomo pogledali, kako dobljene informacije primerno shranimo, da imamo kasneje do njih čim lažji dostop. Konkretno se bomo posvetili procesu postavljanja obratnega indeksa (*ang. inverted index*) ter pripadajočih funkcij za manipuliranje z njim (polnjenje, poizvedbe).

2 Implementacija

2.1 Procesiranje teksta in indeksiranje

Procesiranje `HTML` datotek poteka s pomočjo knjižnice `BeautifulSoup`, `re` ter `nlkt`.

Najprej s knjižnico `BeautifulSoup` odpremo `html` datoteko, odstranimo skripte in vzamemo le vsebino iz razdelka `<body>` ter iz dobljene skrčene vsebine izluščimo tekst.

Dobljeni tekst potem razbijemo s pomočjo klica vgrajene `Python` funkcije `split`, kjer za vsak element dobljenega seznama preverimo, ali vsebuje besedo, ki ni tako imenovan 'stop word'. Če vsebuje primerno besedo, shranimo to besedo pretvorjeno v majhne črke kot 'token' v obliki terke (beseda, indeks v seznamu `text.split()`).

Opomba: Za indeks smo si izbrali indeks glede na tabelo `text.split()`, saj nam to omogoča preprosto rekonstrukcijo izvlečkov ali 'snippetov'.

Za indeksiranje se uporablja SQLite baza. Vsak 'token' pridobljen v prejšnjem koraku (procesiranje teksta) shranimo v bazo, prav tako ime dokumenta v katerem se token nahaja ter indeks njegove pozicije v tekstu. V kolikor se token pojavi na več mestih, se v bazo v stolpec 'indexes' shranijo vsi indeksi pojavitve, ločeni z ','.

2.2 Pridobivanje rezultatov poizvedbe

Z uporabo istih modulov kot pri indeksiranju, poizvedbo najprej sprocesiramo, tako da dobimo 'tokene'. Nato na podlagi teh 'tokenov' poiščemo rezultate poizvedbe.

2.2.1 Obratni indeks

Pri iskanju rezultatov z obratnim indeksom naprej naredimo `SELECT` poizvedbo na bazi, s katero pridobimo vse zapise, katerih beseda je ena izmed 'tokenov'. Nato pridobljene rezultate grupiramo glede na ime dokumenta, tako da za vsak dokument, v katerem se nahaja vsaj eden izmed iskanih 'tokenov', dobimo frekvenco pojavitve 'tokenov' ter indekse pozicij na katerih se nahajajo. Frekvenca je izračunana kot seštevek števila pojavitev vseh 'tokenov' v dokumentu. Na koncu slovar uredimo padajoče po frekvenci in primerno izpišemo rezultate.

2.2.2 Zaporedno pregledovanje dokumentov

Poizvedbo najprej s pomočjo modula na procesiranje teksta spremenimo v 'tokene' in ustvarimo prazen slovar rezultatov, ki bo hranil za vsak dokument, ki vsebuje kakšno primerno besedo terko (frekvenca, indeksi, ime dokumenta). Nato pa se sprehodimo po vseh HTML datotekah v našem naboru, vsako izmed datotek tudi pretvorimo v 'tokene' in na podlagi tega napolnimo naš slovar. Na koncu slovar uredimo padajoče po frekvenci in primerno izpišemo rezultate.

2.2.3 Izpis rezultatov

Pri izpisu rezultatov se izpišejo frekvenca pojavitev iskanih besed, ime dokumenta in 'snippet'. Poizvedba vrne 20 najboljših rezultatov, ki imajo najvišje frekvence iskanih besed. Število vrnjenih rezultatov lahko reguliramo v datoteki `config.py` z določanjem vrednosti parametra `number_of_results`. 'Snippeti' se sestavijo tako, da se za vsak indeks rezultata poizvedbe iz dokumenta preberejo 3 besede, ki se nahajajo pred trenutno obravnavanim 'tokenom' in 3 besede

po njem. V kolikor se nahajata dva 'tokena' eden za drugim, se v 'snippet' vključijo 3 besede po drugem 'tokenu'. Posamezni deli 'snippeta' so ločeni z '...'. V 'snippetih' je izpisanih le prvih 5 odsekov v katerih se pojavijo iskani 'tokeni'.

3 Baza

Baza je sestavljena iz 33.153 različnih besed. Besede, ki se najpogosteje pojavijo v naši zbirki dokumentov so navedene v spodnji tabeli:

Beseda	**Skupno število pojavitev**
podatkov	11054
slovenije	9937
republike	8573
dejavnosti	6101
podatki	5156

Table 1: Tabela petih največkrat pojavljenih besed

Največ pojavitev v enem dokumentu je imela beseda 'proizvodnja', in sicer v dokumentu 'evem.gov.si.371.html' (glej tabelo 2):

Beseda	**Dokument**	**Frekvenca**
proizvodnja	evem.gov.si.371.html	2268
dejavnosti	evem.gov.si.371.html	1512
spada	evem.gov.si.371.html	1338
skupnost	podatki.gov.si.340.html	810
krajevna	podatki.gov.si.340.html	754

Table 2: Tabela prvih petih besed z največ pojavitvami v dokumentu

V tabeli 3 pa lahko vidimo katere strani so vsebovale največ različnih 'tokenov':

Ime dokumenta	**Število različnih 'tokenov'**
evem.gov.si.371.html	12337
podatki.gov.si.340.html	6422
e-prostor.gov.si.57.html	1692
evem.gov.si.398.html	1537
evem.gov.si.651.html	1280

Table 3: Tabela prvih petih strani z največ različnih 'tokenov'

4 Rezultati

Rezultati poizvedb (najboljši trije zadetki oziroma tri najvišje frekvence) so sledeči:

- **“predelovalne dejavnosti”**

Frekvenca	Dokument	Snippet
1516	evem.gov.si.371.html	... iskanje ustrezne šifre dejavnosti /storitve in informacij ... pogojih za opravljanje dejavnosti. V iskalnik vpišite ... 645 od 645 dejavnosti Izpisanih je od ... Izpisanih je od dejavnosti A KMETIJSTVO IN ... pogojih za opravljanje dejavnosti: · Pridelava semenskega ...
77	evem.gov.si.377.html	... Defektolog v zdravstveni dejavnosti Dekan oziroma direktor ... Dietetik v zdravstveni dejavnosti Dimnikar Diplomirana medicinska ... I v zdravstveni dejavnosti Laboratorijski sodelavec II ... II v zdravstveni dejavnosti Laboratorijski tehnik Ladijski ... Logoped v zdravstveni dejavnosti M Magister farmacije ...
42	evem.gov.si.452.html	... e-VEM eVEM Dejavnosti Druge storitvene ... Druge storitvene dejavnosti, drugje nerazvrščene (96.090) ... (96.090) Druge storitvene dejavnosti, drugje nerazvrščene (96.090) ... SKD šifra zajema dejavnosti in storitve, za ... začetek in opravljanje dejavnosti. Predpisi in pogoji: ...

- **“trgovina”**

Frekvenca	Dokument	Snippet
364	evem.gov.si.371.html	... organizacij, gl. 46.110 trgovina na debelo s ... juh, gl. 10.890 trgovina na debelo z ... ipd., gl. 10.890 trgovina na debelo s ... jedmi, gl. 46.380 trgovina na drobno s ... Skladiščenje nevarnih kemikalij Trgovina na debelo z ..... Druga govedoreja Druga trgovina na drobno v ... specializiranih prodajalnah Druga trgovina na drobno v ... nespecializiranih prodajalnah Druga trgovina na drobno v ... z živili Druga trgovina na drobno zunaj ... Nepremičninsko posredovanje Nespecializirana trgovina na debelo Nespecializirana ...
96	evem.gov.si.651.html	... Druga govedoreja Druga trgovina na drobno v ... specializiranih prodajalnah Druga trgovina na drobno v ... nespecializiranih prodajalnah Druga trgovina na drobno v ... z živili Druga trgovina na drobno zunaj ... Nepremičninsko posredovanje Nespecializirana trgovina na debelo Nespecializirana ...
92	evem.gov.si.21.html	... eVEM Področja Trgovina Tu boste našli ... Seznam dejavnosti Druga trgovina na drobno v ... nespecializiranih prodajalnah Druga trgovina na drobno zunaj ... tržnic (47.990) Nespecializirana trgovina na debelo Trgovina ... z drugim blagom Trgovina z zdravili na ...

- **“social services”**

Frekvenca	Dokument	Snippet
5	e-uprava.gov.si.45.html	... Education, culture Labour, retirement Social services, health, death Taxes ... the employment relationship etc.? Social services, health, death How ... I obtain financial social assistance? How do ...
5	e-uprava.gov.si.9.html	... Education, culture Labour, retirement Social services, health, death Taxes ... the employment relationship etc.? Social services, health, death How ... I obtain financial social assistance? How do ...
1	evem.gov.si.661.html	... Records and Related Services (AJPES) and the ...

- **”evidence infrstrukture in načrtov”**

Frekvenca	Dokument	Snippet
51	e-prostor.gov.si.54.html	... kataster gospodarske javne infrastrukture Domov / Zbirke ... kataster gospodarske javne infrastrukture Zbirka vrednotenja nepremičnin ... kataster gospodarske javne infrastrukture Državni prostorski koordinatni ... kataster gospodarske javne infrastrukture Zbirni kataster gospodarske ... kataster gospodarske javne infrastrukture predstavlja temeljno nepremičninsko ...
50	evem.gov.si.371.html	... ministrstvo v okviru evidence oseb, ki imajo ... bil izbrisan iz evidence zbiralcev odpadkov v ... ministrstvo v okviru evidence oseb, ki imajo ... bil izbrisan iz evidence zbiralcev odpadkov v ... ministrstvo v okviru evidence oseb, ki imajo ...
15	evem.gov.si.203.html	... omrežij in pripadajoče infrastrukture na področju elektronskih ... omrežij in pripadajoče infrastrukture na področju elektronskih ... omrežij in pripadajoče infrastrukture v skladu s ... omrežij in pripadajoče infrastrukture za potrebe varnosti, ... omrežij in pripadajoče infrastrukture na nepremičninah v ...

- **"poceni nepremičnine"**

Frekvenca	Dokument	Snippet
15	podatki.gov.si.340.html	... storitve, d.o.o. AKM NEPREMIČNINE, družba za upravljanje ... d.o.o. ARG - Nepremičnine d.o.o. ARGOLINA investicijski ... družba za inženiring, nepremičnine, urbanizem in energetiko, ... merilni laboratorij in nepremičnine, d.d. ELEKTROTEHNIŠKO-RAČUNALNIŠKA STROKOVNA ... - NOTARKA FITERA, nepremičnine in trgovina d.o.o. ...
12	e-prostor.gov.si.12.html	... prostorskih podatkov / Nepremičnine / Register nepremičnin ... Evidenca trga nepremičnin Nepremičnine Zemljiški kataster Kataster ... Gradiva Podatki Statistike Nepremičnine namenjene opravljanju dejavnosti ... Vsi deli ene nepremičnine (zemljišča in deli ... so verjetno lastnik nepremičnine. Predvsem to velja ...
9	e-prostor.gov.si.51.html	... Evidenca trga nepremičnin Nepremičnine Topografski in kartografski ... prave vrednosti moje nepremičnine? Posnetek predavanja mag. ... kjer imajo podobne nepremičnine približno enako vrednost, ... določajo vpliv lokacije nepremičnine na vrednost. Lokacijo ... izraža vrednost referenčne nepremičnine v posamezni vrednostni ...

- **"registracija"**

Frekvenca	Dokument	Snippet
35	evem.gov.si.68.html	... položaja točke SPOT registracija Vsebina je v ... subjektom - SPOT registracija . Položaj točke ... Položaj točke SPOT registracija se pridobi na ... položaj točke SPOT registracija Položaj točke za ... subjektom – SPOT registracija lahko pridobi: upravna ...
15	evem.gov.si.67.html	... Točke - SPOT Registracija Poiščite najbližjo točko ... najbližjo točko SPOT registracija (VEM točka) . ... na točkah SPOT Registracija vam bodo pomagali ... Storitve točk SPOT Registracija izvajajo: AJPES, Upravne ... Storitve točke SPOT Registracija (VEM točke) so ...
9	evem.gov.si.653.html	... imeniku ZAPS - revidenti Registracija Registracija čolna Registracija fitofarmacevtskih ... fitofarmacevtskih sredstev (FFS) Registracija in označevanje stojišča ... in označevanje stojišča Registracija živilskih obratov Ribolovna ... kemikalij Status in registracija Status in registracija ... registracija Status in registracija Status in registracija ...

4.1 Hitrosti poizvedb

Poizvedba	Čas poizvedbe (baza)	Čas poizvedbe (zaporedno)
predelovalne dejavnosti	6,0671 s	71,2128 s
trgovina	6,0069 s	69,8861 s
social services	1,4969 s	66,0935 s
evidence infrastrukture in načrtov	6,3728 s	67,6538 s
poceni nepremičnine	6,0568 s	71,5132 s
registracija	4,5208 s	67,6229 s