Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Mihael Švigelj, Lovro Vražič in Luka Zakšek

Indexing

Domača naloga pri predmetu Iskanje in ekstrakcija podatkov s spleta

MENTOR: prof. dr. Marko Bajec

V nalogi smo razvili ekstrakcijo teksta iz podanih spletnih strani, iz katerega smo zgradili inverted index v SQLite, s pomočjo katerega lahko hitro iščemo najbolj pomembne spletne strani glede na podane ključne besede.

1. Uvod

Nalogo smo implementirali v jeziku Python. Pri implementaciji smo uporabili knjižnice: sqlite3, Beautiful Soup in nltk.

2. Predprocesiranje in indeksiranje

Text iz podanih spletnih strani ekstrahiramo z uporabo Beautiful Soup. Dobljeni tekst normaliziramo v male črke, potem pa ga z uporabo nltk knjižnice razdelimo v tokene iz katerih pa še odstranimo tako imenovane “stopword”-e. Vse unikatne besede v tako dobljenih tokenih vstavimo v podatkovno bazo, vključno z imenom dokumenta, številom pojavitev in indeksi pojavitev znotraj dokumenta.

3. Iskanje z inverted index

Vnos za iskanje najprej sprocesiramo na enak način, kot tekste pred indeksiranjem, to je - normaliziramo v male črke in z nltk knjižnico tokeniziramo besede in odstranimo stopworde. Nato v bazi iščemo vse vnose s podanimi besedami. Dobljene dokumente uredimo po frekvenci najdenih besed. Za izpis dokumente ponovno odpremo in ponovno sprocesiramo na enak način kot pri indeksiranju, da indeksi v bazi kažejo na iste besede v dokumentu. Izpišemo še 3 besede pred in po iskani besedi. Tak način ni idealen, izpis je brez stopwordov, brez ločil in velikih začetnic.

4. Naivno iskanje

Iskanje je zelo preprosto. Najprej na enak način kot pri indeksiranju sprocesiramo vnos za iskanje. Potem vsako datoteko posebej sprocesiramo in preštejemo število ponovitev katere

koli izmed besed v iskalnem nizu. Ko so pregledane vse datoteke se v vrstnem redu po frekvenci najdenih besed izpišejo rezultati.

5. Opis podatkovne baze

Posting ima 403070 zapisov. IndexWord ima 49121 zapisov.

6. Primeri poizvedovanj

- “predelovalne dejavnosti”

Frequency	Document	Snippet
1284	data/evem.gov.si/evem.gov.si.371.html	... iskanje ustrezne šifre dejavnosti /storitve informacij pogojih ... informacij pogojih opravljanje dejavnosti. iskalnik vpišite ...
75	data/evem.gov.si/evem.gov.si.377.html	... straže defektolog zdravstveni dejavnosti dekan direktor delavec ... detektiv dietetik zdravstveni dejavnosti dimnikar diplomirana medicinska ...
40	data/podatki.gov.si/podatki.gov.si.340.html	... - nosilec dopolnilne dejavnosti kmetiji bregar miro ... šport center interesnih dejavnosti ptuj center judovske ...
39	data/evem.gov.si/evem.gov.si.452.html	... ... 96.090) / dejavnosti / evem republika ... e-vem evem›dejavnosti›druge storitvene dejavnosti, drugje nerazvrščene ...
31	data/evem.gov.si/evem.gov.si.653.html	... licenca dovoljenje opravljanje dejavnosti specializirane prodajalne zdravili ... izvajanje radijske televizijske dejavnosti dovoljenje izvajanje sevalne ...

- “trgovina”

Frequency	Document	Snippet
364	data/evem.gov.si/evem.gov.si.371.html	... gl. 46.110 trgovina debelo kmetijskimi pridelki ... gl. 10.890 trgovina debelo mesnimi izdelki ...
94	data/evem.gov.si/evem.gov.si.651.html	... trgu dozimetrija govedoreja trgovina drobno specializiranih prodajalnah ... drobno specializiranih prodajalnah trgovina drobno nespecializiranih prodajalnah ...

92	data/evem.gov.si/evem.gov.si.21.html	... zapiram e-vem evem›področja trgovina našli informacije pogojih ... . seznam dejavnosti trgovina drobno nespecializiranih prodajalnah ...
82	data/podatki.gov.si/podatki.gov.si.340.html	... . dent, trgovina storitve, d.o.o ... . adria investicije trgovina, posredništvo, ...
13	data/evem.gov.si/evem.gov.si.623.html	... izdelki široke porabe trgovina debelo izdelki široke ... porabe spada : trgovina debelo lesenimi, ...

- “social services”

Frequency	Document	Snippet
3	data/e-uprava.gov.si/e-uprava.gov.si.45.html	... labour, retirement social services, health ... employment relationship ? social services, health ... can obtain financial social assistance ? how
3	data/e-uprava.gov.si/e-uprava.gov.si.9.html	... labour, retirement social services, health ... employment relationship ? social services, health ... can obtain financial social assistance ? how
2	data/e-uprava.gov.si/e-uprava.gov.si.45.html	... , retirement social services, health, ... relationship ? social services, health,
2	data/e-uprava.gov.si/e-uprava.gov.si.9.html	... records and related services (ajpes)
1	data/evem.gov.si/evem.gov.si.661.html	... procese : • test popolnosti – preverite

- “vložite”

Frequency	Document	Snippet
3	data/evem.gov.si/evem.gov.si.406.html	... poslovodna oseba) vložite obrazcu m-2. ... m-2. odjavo vložite, izpolnjeni pogoji ... rubrik. delodajalec vložite odjavo obveznih socialnih
1	data/evem.gov.si/evem.gov.si.23.html	... datumom vpisa, vložite želenim datumom vpisa

1	data/evem.gov.si/evem.gov.si.366.html	... statusa invalidskega podjetja vložite ministrstvo delo družino
1	data/evem.gov.si/evem.gov.si.402.html	... poklicne bolezni prijava/odjava vložite obrazcu m12 :
1	data/evem.gov.si/evem.gov.si.48.html	... zavarovanj obrazcem m-2 vložite dneh dnevom izbrisa

- “slovenska vlada”

Frequency	Document	Snippet
45	data/podatki.gov.si/podatki.gov.si.340.html	... 2. osnovna šola slovenska bistrica 2tdk, ... center socialno delo slovenska bistrica center socialno ... gradec glasbena šola slovenska bistrica glasbena šola ... kabelsko televizijo informiranje slovenska bistrica javni zavod ...
27	data/podatki.gov.si/podatki.gov.si.350.html	... podrobnosti organizacija : vlada republike slovenije statistični ... urad republike slovenije vlada republike slovenije statistični ... mesečno 828 ogledov vlada republike slovenije statistični ... metapodatki, vir vlada republike slovenije statistični ...
22	data/podatki.gov.si/podatki.gov.si.364.html	... 3) organizacija vlada republike slovenije statistični ... (27) vlada republike slovenije služba ... letno 13 ogledov vlada republike slovenije statistični ... metapodatki, vir vlada republike slovenije statistični ... letno 13 ogledov vlada republike slovenije statistični ... metapodatki, vir vlada republike slovenije statistični ...
22	data/podatki.gov.si/podatki.gov.si.88.html	... 3) organizacija vlada republike slovenije statistični ... (27) vlada republike slovenije služba ... letno 10 ogledov vlada republike slovenije statistični ... metapodatki, vir vlada republike slovenije statistični ... letno 10 ogledov vlada republike slovenije statistični ... metapodatki, vir vlada republike slovenije statistični ...

21	data/podatki.gov.si/podatki.gov.si.114.html	... 39) organizacija vlada republike slovenije statistični ... večletno 109 ogledov vlada republike slovenije statistični ... metapodatki, vir vlada republike slovenije statistični ... letno 52 ogledov vlada republike slovenije statistični ...

- “ministrstvo za finance”

Frequency	Document	Snippet
589	data/evem.gov.si/evem.gov.si.371.html	... rastlin vir : ministrstvo kmetijstvo, gozdarstvo ... živil vir : ministrstvo kmetijstvo, gozdarstvo ... živil vir : ministrstvo kmetijstvo, gozdarstvo ...
48	data/podatki.gov.si/podatki.gov.si.340.html	... bojan - izvršitelj ministrstvo delo, družino ... republike slovenije delo ministrstvo finance ministrstvo finance ... delo ministrstvo finance ministrstvo finance finančna uprava ... uprava republike slovenije ministrstvo finance uprava republike ...
45	data/podatki.gov.si/podatki.gov.si.348.html	... republika slovenija, ministrstvo delo, družino ... (61) ministrstvo okolje prostor, ... (58) ministrstvo notranje zadeve (... (58) ministrstvo javno upravo (... (42) ministrstvo infrastrukturo (39 ... (39) ministrstvo kmetijstvo, gozdarstvo ...
42	data/podatki.gov.si/podatki.gov.si.31.html	... republika slovenija, ministrstvo delo, družino ... (61) ministrstvo okolje prostor, ... (58) ministrstvo notranje zadeve (... (58) ministrstvo javno upravo (... (42) ministrstvo kmetijstvo, gozdarstvo ... (39) ministrstvo infrastrukturo (39 ... (39) ministrstvo okolje prostor, ... (33) ministrstvo okolje prostor, ... (30) ministrstvo finance (30 ...
30	data/podatki.gov.si/podatki.gov.si.214.html	... , gozdarstvo prehrana finance davki gospodarstvo energetika ... , gozdarstvo prehrana finance davki gospodarstvo energetika ... organizacija : ministrstvo finance ministrstvo finance vključi ...

		ministrstvo finance ministrstvo finance vključi zbirke hčerinskih ...

Rezultati poizvedb so pravilni, v smislu, da poizvedbe res vrnejo strani, ki vsebujejo iskane besede, vendar pa rezultati niso optimalni. Želeli bi, da imajo dokumenti, v katerih se najdejo natančni pari besed ali celi stavki višjo težo, ne zgolj da se seštejejo pojavitve katere koli izmed besed. Prav tako bi idealno inverted-index upošteval tudi dolžino samega dokumenta in pa splošno pogostost iskanih besed. Koristna bi bila tudi implementacija lematizacije besed.

7. Zaključek

Implementirali smo inverted-index na SQL bazi. Poizvedovanje deluje solidno. Želeli bi si hitrejše poizvedovanje, večjo splošnost z lematizacijo besed, lepše povzetke rezultatov in boljše vrednotenje relevantnih zadetkov. Implementirali smo tudi naivno metodo preiskovanja, ki dokazuje, da je kljub navidezni počasnosti v primerjavi z state-of-the-art iskalniki, implementirani inverted-index še vedno mnogo boljša rešitev.