Blaž Štempelj in Denis Grabljevec

Seminarska naloga: Indeksiranje in pridobitev podatkov

Uvod

Inverzni indeks[1] je najbolj uporabljena podatkovna struktura v sistemih za ekstrakcijo podatkov. Omogoča nam hitro iskanje po tekstu, podaljša pa čas vstavljanja dokumenta v bazo. Razlog je ta, da mora prej zgraditi nov indeks, ki vsebuje tudi besede iz novega dokumenta. V seminarski nalogi sva torej iz kolekcije spletnih strani iz štirih različnih domen zgradila indeks in ga testirala s parimi poizvedbami.

Implementacija

Za implementacijo inverznega indeksa sva uporabila programski jezik Python[2] in podatkovno bazo PostgreSQL[3]. Napisala sva program, ki je s pomočjo knjižnice BeautifulSoup pridobil vsebino iz več kot tisoč spletnih strani iz štirih različnih domen. Pridobitvi vsebine je sledila njena obdelava, kjer sva si pomagala še s knjižnico NLTK[4]. Uporabila sva NLTK funkcijo *word_tokenizer*, ki besedilo razbije na posamezne lematizirane besede. Nato vsa odstranila še tako imenova *stopwords* besede, ki niso pomembne za razumevanje samega besedila. Iz pridobljenega prečiščenega seznama sva si nato za vsako besedo zapomnila *spletno stran* kjer se pojavi, *število ponovitev* in *pozicijo v besedilu*, npr. `<stran1, 3, [1, 4, 15]>`. Pridobljene besede, ponovitve besede in pozicijo sva shranila v dve tabeli v podatkovni bazi. Tabela *IndexWord* hrani vse najdene besede (brez dvojnikov), tabela *Posting* pa hrani spletno stran kjer se beseda pojavi, število ponovitev in pozicije v besedilu na katerih se beseda pojavi.

Pridobivanje podatkov sva se lotila tako, da sva iskane besede razdelila glede na presledke, ter nato v bazi preko indexa poiskala v katerih dokumentih in kolikokrat se beseda pojavlja. Izpisala sva tudi vse indexe, kje v besedilu se beseda nahaja (merge indexov v queryju). S relativno enostavnim queryjom, hitro dobimo podatke na katerem mestu v datoteki se nahaja določena beseda.
Kot zadnje sva implementirala tudi algoritem izpis treh besed pred in po iskani besedi (snippet).

Rezultati

Iskanje dokumentov in izgradnjo snippet-ov iz iskalnih besed, je po zaslugi zgornje implementacije zelo hitra operacija.

Iskalni nizi:

1. **predelovalne dejavnosti**
   75 evem.gov.si/evem.gov.si.377.html: Defektolog v zdravstveni dejavnosti Dekan oziroma direktor ... Dietetik v zdravstveni dejavnosti Dimnikar Diplomirana medicinska ... začetek in opravljanje dejavnosti. Predpisi in pogoji …

   37 evem.gov.si/evem.gov.si.452.html: za opravljanje dopolnilne dejavnosti na kmetiji Dovoljenje ... za opravljanje strateške dejavnosti Dovoljenje za opravljanje ... dovoljenje za opravljanje dejavnosti proizvodnje nevarnih kemikalij ...

   18 evem.gov.si/evem.gov.si.28.html za opravljanje gospodarske dejavnosti. Lastnosti zasebnega zavoda … niso za posamezne dejavnosti ali posamezne vrste … iz opravljanja nepridobitne dejavnosti se ne obdavči

   4 podatki.gov.si/podatki.gov.si.228.html glede na področje dejavnosti in vrednost njihove ... glede na področje dejavnosti in vrednost njihove ... glede na področje dejavnosti in vrednost njihove …

2. **trgovina**
   364 evem.gov.si/evem.gov.si.371.html Sem ne spada: trgovina na debelo za … Sem ne spada: trgovina na debelo za … kovinskimi izdelki, ključavnicami trgovina na debelo z …

   9 evem.gov.si/evem.gov.si.327.html napravami in opremo Trgovina na debelo z ... opremo Sem spada: trgovina na debelo s ... koles in koles trgovina na debelo z ... z industrijskimi roboti trgovina na debelo z …

   7 evem.gov.si/evem.gov.si.316.html in rudami(46.720) Trgovina na debelo s ... 46.720) Sem spada: trgovina na debelo z ... debelo z rudami trgovina na debelo z ... v primarni obliki trgovina na debelo s

   2 evem.gov.si/evem.gov.si.312.html s fitofarmacevtskimi sredstvi Trgovina na debelo s ... Pridobitev dovoljenja Področje: Trgovina SKD številka: 46.750 .

   2 evem.gov.si/evem.gov.si.326.html s tobačnimi izdelki Trgovina na debelo s ... ni možno. Področje: Trgovina SKD številka: 46.350 .

   2 evem.gov.si/evem.gov.si.338.html in tobačnimi izdelki Trgovina na drobno na ... ni možno. Področje: Trgovina SKD številka: 47.810 .

   2 evem.gov.si/evem.gov.si.347.html s tobačnimi izdelki Trgovina na drobno v ... ni možno. Področje: Trgovina SKD številka: 47.260 .

3. **social services**

   5 e-uprava.gov.si/e-uprava.gov.si.45.html Labour, retirement Social services, health, death Taxes ... relationship etc.? Social services, health, death How ... culture Labour, retirement Social services, health, death ... employment relationship etc.? Social services, health, death ... I obtain financial social assistance? How do

   5 e-uprava.gov.si/e-uprava.gov.si.9.html Labour, retirement Social services, health, death Taxes ... relationship etc.? Social services, health, death How ... culture Labour, retirement Social services, health, death ... employment relationship etc.? Social services, health, death ... I obtain financial social assistance? How do

   1 evem.gov.si/evem.gov.si.661.html Records and Related Services(AJPES) and the

   1 podatki.gov.si/podatki.gov.si.340.html recreation and spa services ltd. TERME MARIBOR .

4. **ministrstvo za izobraževanje**

   102 podatki.gov.si/podatki.gov.si.340.html ... SLOVENIJE ZA POMORSTVO MINISTRSTVO ZA IZOBRAŽEVANJE, ZNANOST ... ZNANOST IN ŠPORT MINISTRSTVO ZA IZOBRAŽEVANJE, ZNANOST ... tem zakonu. Vir: Ministrstvo za izobraževanje, znanost … farmacevtske namene; Vir: Ministrstvo za gospodarski razvoj ...

   25 podatki.gov.si/podatki.gov.si.119.html REPUBLIKE SLOVENIJE(416) MINISTRSTVO ZA IZOBRAŽEVANJE, ZNANOST ... IN ŠPORT(3) MINISTRSTVO ZA KULTURO(1 ... objektov 77 ogledov MINISTRSTVO ZA IZOBRAŽEVANJE, ZNANOST ... dediščine 216 ogledov MINISTRSTVO ZA KULTURO- Register ... programov 93 ogledov MINISTRSTVO ZA IZOBRAŽEVANJE, ZNANOST ...

   18 podatki.gov.si/podatki.gov.si.404.html REPUBLIKE SLOVENIJE(416) MINISTRSTVO ZA IZOBRAŽEVANJE, ZNANOST ... IN ŠPORT(3) MINISTRSTVO ZA KULTURO(

   8 evem.gov.si/evem.gov.si.452.html telesa) ; Vir: Ministrstvo za gospodarski razvoj ... najemnika storitev. Vir: Ministrstvo za gospodarski razvoj ... njih pa se ministrstvo, pristojno za okolje ... največ petih let.Vir: Ministrstvo za okolje in ...

   3 podatki.gov.si/podatki.gov.si.164.html Izobraževanje, kultura in šport Sociala in zaposlovanje ... Izobraževanje, kultura in šport Sociala in zaposlovanje ... Izobraževanje, kultura in šport Sociala in zaposlovanje

5. **Vlada republike Slovenije**

95 podatki.gov.si/podatki.gov.si.340.html za raziskovalno dejavnost Republike Slovenije JAVNA AGENCIJA ... ZA ŽELEZNIŠKI PROMET REPUBLIKE SLOVENIJE JAVNA RAZSVETLJAVA …

66 evem.gov.si/evem.gov.si.371.html državljan oziroma državljanka Republike Slovenije ali tuji ... oziroma polnoletna državljanka Republike Slovenije, ki ima ... ki na teritoriju Republike Slovenije proizvajajo nevarne ... ki na teritoriju Republike Slovenije proizvajajo nevarne ... ki na teritoriju Republike Slovenije proizvajajo nevarne ... ki na teritoriju Republike Slovenije proizvajajo nevarne ... ki na teritoriju Republike Slovenije proizvajajo nevarne ... ki na teritoriju Republike Slovenije proizvajajo nevarne ... ki na teritoriju Republike Slovenije proizvajajo nevarne ... ki na teritoriju Republike Slovenije proizvajajo nevarne ... ki na teritoriju Republike Slovenije proizvajajo nevarne ... P2, na območju Republike Slovenije pristojni organ ... P2, na območju Republike Slovenije pristojni organ ... ki na teritoriju Republike Slovenije proizvajajo nevarne ...

13 podatki.gov.si/podatki.gov.si.231.html Organizacije Podrobnosti VLADA REPUBLIKE SLOVENIJE STATISTIČNI URAD ... SLOVENIJE STATISTIČNI URAD REPUBLIKE SLOVENIJE Objavil: VLADA ... SLOVENIJE Objavil: VLADA REPUBLIKE SLOVENIJE STATISTIČNI URAD

6 podatki.gov.si/podatki.gov.si.220.html 1) Organizacija VLADA REPUBLIKE SLOVENIJE STATISTIČNI URAD ... SLOVENIJE STATISTIČNI URAD REPUBLIKE SLOVENIJE(1) Ocena ... 31 ogledov VLADA REPUBLIKE SLOVENIJE STATISTIČNI URAD ... SLOVENIJE STATISTIČNI URAD REPUBLIKE SLOVENIJE- V tej ... vir je VLADA REPUBLIKE SLOVENIJE STATISTIČNI URAD ... SLOVENIJE STATISTIČNI URAD REPUBLIKE SLOVENIJE in ustrezajo

6. **Univerza v ljubljani**

53 podatki.gov.si/podatki.gov.si.340.html fakultete Univerze v Ljubljani CENTER REPUBLIKE SLOVENIJE ... KNJIŽNICA UNIVERZE V LJUBLJANI CENTRO ITALIANO DI ... SOCIALNO SODIŠČE V LJUBLJANI DELOVNO SODIŠČE CELJE ... INŠTITUT UNIVERZE V LJUBLJANI INSTITUT INFORMACIJSKIH ZNANOSTI ... INŠTITUT UNIVERZE V LJUBLJANI FAKULTETE ZA GRADBENIŠTVO ... Pravni fakulteti v Ljubljani INŠTITUT ZA EKONOMSKA

4 podatki.gov.si/podatki.gov.si.542.html v Poligonu v Ljubljani Pod Črto je ... za upravo v Ljubljani se je 4 ... v Poligonu v Ljubljani Pod Črto je ... Gospodarskem razstavišču v Ljubljani. Nadaljujte z branjem

1 podatki.gov.si/podatki.gov.si.291.html se je v Ljubljani odvijal hackaton

1 podatki.gov.si/podatki.gov.si.534.html Študentje Univerze v Ljubljani, Fakultete za računalništvo

Zaključek

Implementacija inverznega indeksa s pomočjo podatkovne baze v splošnem ni zahtevna naloga. Potrebno pa je biti zelo pozoren pri detajlih. Besedilo, ki ga ekstrahiramo iz spletnih strani mora biti prečiščeno. To pomeni, da odstranimo vse značke html in vso dodatno JavaScript kodo. Poleg tega se mora indeks ujemati z ekstrahiranim besedilom. Le-to ujemanje se lahko v trenutku podre, če odstranimo nepomembne besede iz beseda, indeks pa njihov obstoj ne upošteva. V besedilu, ki ga dobi uporabnik po vnosu poizvedbe morajo te besede seveda biti prisotne, drugače lahko stavki povsem izgubijo svoj pomen ali pa postanejo popolnoma neberljivi.

Literatura

[1] Inverted Index. (maj 2019) [Online]. Dosegljivo: https://en.wikipedia.org/wiki/Inverted_index
[2] Python. (maj 2019) [Online]. Dosegljivo: https://www.python.org/
[3] PostgreSQL. (maj 2019) [Online]. Dosegljivo: https://www.postgresql.org/
[4] NLTK. (maj 2019) [Online]. Dosegljivo: http://www.nltk.org/