Indeksiranje in iskanje dokumentov

Poročilo seminarske naloge

Adam Prestor, Lojze Žust
ap2408@student.uni-lj.si, lojze.zust@student.uni-lj.si

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

1 Uvod

V prvi seminarski nalogi smo se osredotočili na preiskovanje spleta, v drugi smo naredili ekstrakcijo podatkov z zajetega spleta, v tej nalogi pa smo se osredotočili na poizvedovanje. Podatke, ki smo jih zajeli s spleta smo v prvem delu obdelali z metodami obdelave naravnega jezika, kot sta tokenizacija in lematizacija ter zgradili invertni indeks. V drugem delu pa smo indeks uporabili za rangiranje dokumentov in vračanju najbolj relevantnih. Poleg tega smo primerjali hitrost poizvedbe s hitrostjo poizvedbe, če ne bi uporabili indeksa temveč smo uporabili surovo vsebino strani.

2 Grajenje indeksa

Prvi del metode predstavlja razčlenjevanje in obdelava vsebine spletnih strani in gradnja invertnega indeksa. Za razčlenjevanje HTML vsebine smo uporabili knjižnico `BeautifulSoup`. Iz HTML vsebine najprej odstranimo vse `script` in `style` značke, ki predstavljajo programsko kodo in stile strani, ter ne vsebujejo same vsebine. Nato pridobimo vso tekstovno vsebino (brez značk) preostanka strani z uporabo metode `soup.get_text()`. Nad pridobljenim besedilom nato izvedemo predprocesiranje, ki iz besedila pridobi tokene. Enak postopek se izvede kasneje nad besedilom pozvedbe.

2.1 Predprocesiranje

Predprocesiranje poteka v več korakih. V prvem koraku izvedemo osnovno tokenizacijo z metodo `word_tokenize` iz knjižnjice *nltk*. Nato izvedemo lematizacijo, za kar uporabimo slovensko različico lematizatorja *Lemmagen*. S tem korakom želimo omogočiti, da poizvedbe najdejo tudi rezultate z besedami v slovnično drugačnih oblikah (npr. predelovalne dejavnosti -> predelovalna dejavnost). Slabost trenutnega pristopa je, da deluje le za slovenski jezik, kar bi se dalo odpraviti z nadgradnjo v prihodnosti. V naših enostavnih poskusih je večina angleških besed po lematizaciji ostala nespremenjena, tako da negativen učinek za angleške strani ni prevelik.

Po lematizaciji se vsi tokeni pretvorijo v male črke, nato pa se odstranijo še stop besede iz slovarja. Za stop besede smo vzeli slovar iz navodil naloge, s

tem da smo ga morali rahlo prilagoditi, da je deloval. Pri nas je *nltk* namreč za slovar slovenskih besed imel drugačno ime, kot v skripti iz navodil, ter ga ta ni uspela najti. V tem koraku z enostavnim regex vzorcem odstranimo tudi vse tokene sestavljene zgolj iz ne-alfanumeričnih znakov (npr. ločila ipd.). Izhod predprocesiranja se nato uporabi za grajenje indeksa. Tega zgradimo tako, da gremo čez vse tokene in si beležimo število pojavitev in indekse posameznih besed. Te informacije potem vnesemo v podatkovno bazo.

3 Podatkovna baza

Podatkovno bazo smo naredili po shemi iz navodil. Vsebuje dve tabeli, *IndexWord* in *Posting*.

V *IndexWord* shranjujemo unikatne vrednosti, ki se pojavljajo v indeksu. Teh besed je v našem primeru 34697.

Tabela *Posting* pa vsebuje več informacij, saj je v njej poleg besede tudi dokument v katerem se nahaja, njena frekvenca in položaji besede v dokumentu. Ta tabela ima 325600 vrstic. Najbolj pogosti pari besed in dokumentov so: <proizvodnja><evem.gov.si/evem.gov.si.371.html>, <dejavnost><evem.gov.si/evem.gov-.si.371.html>, <spadati><evem.gov.si/evem.gov.si.371.htm> in <gti><evem.gov-.si/evem.gov.si.371.html>. Največja frekvenca je 5238 in sicer za prvi par. Največkrat se v bazi pojavi stran <evem.gov.si/evem.gov.si.371.html>, pojavi se sicer 8169 krat. Med besedami pa je najbolj pogosta <uporaba>, ki se pojavi 1401 krat.

3.1 Poizvedovanje

Poizvedovanje smo naredili s pomočjo invertnega indeksa ter brez, se pravi smo uporabili surove podatke zajete s strani. V obeh primerih začnemo s poizvedbo, ki jo obdelamo enako, kot smo obdelali vsebino strani. Metode so predstavljene v poglavju 2. Tako dobimo enake žetone, kot jih dobimo s strani. S temi ženoti potem vstopimo v poizvedbo.

3.2 Poizvedovanje z indeksom

Ko poizvedujemo s pomočjo indeksa, najprej poiščemo vse strani, ki so shranjeni v bazi. Potem se sprehodimo po vsaki strani, ter iščemo vse pare žetonov in strani ter si zapomnimo njihovo frekvenco in položaje. Če je žetonov več, frekvence seštejemo in si zapomnimo njihove položaje. Te podatke dodamo v seznam skupaj z imenom strani. Ko se sprehodimo čez vse, seznam uredimo padajoče in izberemo le n najvišjih strani. Za te strani, s pomočjo pozicij žetonov, generiramo snipete.

3.3 Surovo poizvedovanje

Surovo poizvedovanje podobno kot s pomočjo indeksa. Razlika je v tem, da ko stran izberemo, moramo njeno vsebino vsakič znova obdelati z metodami

opisanimi v poglavju 2, kar je zamudno. Tako pretvorimo vsebino strani v žetone. Potem se sprehodimo skozi vse žetone, če žeton sovpada z žetonom iz poizvedbe, potem povečamo frekvenco in si zapomnimo njegovo pozicijo. Za vsako stran si zapomnimo: frekvenco, stran in pozicijo žetonov. To dodamo v seznam, ki ga na koncu padajoče uredimo po frekvenci ter izberemo le n najvišjih strani, za katere generiramo še snipete.

3.4 Generiranje snipetov

Za generiranje snipetov potrebujemo pozicije žetonov. Vsako pozicijo potem razširimo na okolico 3 besed pred in po ter odstranimo duplikate. Potem poiščemo zaporedne pozicije in jih združimo v enoto. Te enote bodo predstavljale posamezne odseke ločene s tropičjem. Ker so snipeti lahko zelo dolgi, zato smo vzeli le 250 znakov in tako nekako skrajšali v predstavljivo obliko.

3.5 Primerjava Rezultatov

Metode smo poganjali z naslednjimi poizvedbami:

- predelovalne dejavnosti
- trgovina
- social services
- socialno delo
- varovanje človekovih pravic
- seja državnega zbora

Za vse poizvedbe smo vzeli 6 najvišje rangiranih strani. Rezultati vrnjenih strani se ne razlikujejo, glede na metodo poizvedovanja. Kar pa se razlikuje je čas, ki je potreben za eno in drugo metodo. Poizvedovanje s pomočjo indeksa je 120 krat hitrejše od surovega poizvedovanja.

Vsi rezultati so predstavljeni na straneh, ki sledijo zaključku.

4 Zaključek

V nalogi smo implementirali enostvano različico poizvedovanja po inverted indeksu in po goli vsebini. Iz rezultatov je jasno razvidno, da je poizvedovanje z indeksom mnogo hitrejšo. V našem primeru za faktor 120. Ta faktor bi bil verjento še večji, če bi imeli več strani oz. dokumentov v naši zbirki.

Results for a query: “predelovalne dejavnosti”
Results found in 0.35s
Raw search results found in 63.05s

Frequency	Document	Snippet
1885	evem.gov.si/evem.gov.si.371.html	iskanje ustrezne šifre dejavnosti /storitve in informacij ... pogojih za opravljanje dejavnosti . V iskalnik ... 645 od 645 dejavnosti Izpisanih je od dejavnosti A KMETIJSTVO IN ... pogojih za opravljanje dejavnosti : · Pridelava ... pogojih za oprav
133	podatki.gov.si/podatki.gov.si.340.html	A-DENT ZOBJE zobozdravstvena dejavnost d.o.o . 2leva ... ADAMLJE NEVIJA - DEJAVNOST PATRONAŽE IN ZDRAVSTVENE ... podjetje za letališko dejavnost , gostinstvo in ... KOROŠEC ANDREJA - DEJAVNOST ZOBOZDRAVSTVA ODRASLIH AMBROŽIČ ... , splošna zdravstven
128	evem.gov.si/evem.gov.si.32.html	, ki opravljajo dejavnost 31 Mar Predložitev ... , ki opravljajo dejavnost 21 Mar Večstranski ... , ki opravljajo dejavnost 20 Mar Banka ... , ki opravljajo dejavnost 18 Mar Rok ... , ki opravljajo dejavnost 18 Mar Poročanje ... , ki opravljajo dejav
79	evem.gov.si/evem.gov.si.377.html	Defektolog v zdravstveni dejavnosti Dekan oziroma direktor ... Dietetik v zdravstveni dejavnosti Dimnikar Diplomirana medicinska ... delavci v sevalnih dejavnostih Izvajalec elektroribolova Izvajalec ... I v zdravstveni dejavnosti Laboratorijski sode
55	evem.gov.si/evem.gov.si.452.html	Druge storitvene dejavnosti , drugje nerazvrščene ... 96.090) / Dejavnosti / eVEM Republika ... e-VEM eVEM › Dejavnosti › Druge storitvene dejavnosti , drugje nerazvrščene ...) Druge storitvene dejavnosti , drugje nerazvrščene ... SKD šifra zajema
48	evem.gov.si/evem.gov.si.398.html	Opravljamo samo oproščeno dejavnost , od katere ... mesecev opravljala ekonomsko dejavnost ali ne . ... začnemo dejansko opravljati dejavnost ? Gospodarska družba ... usmerjene na opravljanje dejavnosti (npr . ... da namerava opravljati dejavnost,

Results for a query: “seja državnega zbora”
Results found in 0.33s
Raw search results found in 61.05s

Frequency	Document	Snippet
72	e-prostor.gov.si/e-prostor.gov.si.13.html	E-prostor - Državna meja Ta stran ... O portalu Kontakt Državne ustanove – – ... Ministrstva Predsednik RS Državni zbor Državni svet Ustavno sodišče ... Zbirke prostorskih podatkov Državna meja Domov / ... / Nepremičnine / Državna meja Zbirka vredn
49	podatki.gov.si/podatki.gov.si.106.html	DRŽAVNI ZBOR REPUBLIKE SLOVENIJE - ... Organizacija : DRŽAVNI ZBOR REPUBLIKE SLOVENIJE DRŽAVNI ZBOR REPUBLIKE SLOVENIJE Vključi ... Datasets Priljubljene zbirke Državni organi Javni sektor ... 78 ogledov DRŽAVNI ZBOR REPUBLIKE SLOVENIJE - ... branjem
46	e-prostor.gov.si/e-prostor.gov.si.17.html	- Zbirka podatkov državnih geodetskih točk Ta ... O portalu Kontakt Državne ustanove – – ... Ministrstva Predsednik RS Državni zbor Državni svet Ustavno sodišče ... podatkov Zbirka podatkov državnih geodetskih točk Domov ... prostorskih podatkov /
43	podatki.gov.si/podatki.gov.si.23.html	Organizacija DRŽAVNI ZBOR REPUBLIKE SLOVENIJE Ocena ... Zadnji posodobitvi Go Državni organi Javni sektor ... 78 ogledov DRŽAVNI ZBOR REPUBLIKE SLOVENIJE - ... branjem ... XML Državni organi Javni sektor ... 27 ogledov DRŽAVNI ZBOR REPUBLIKE SLOVEN
43	podatki.gov.si/podatki.gov.si.373.html	Organizacija DRŽAVNI ZBOR REPUBLIKE SLOVENIJE Ocena ... Zadnji posodobitvi Go Državni organi Javni sektor ... 78 ogledov DRŽAVNI ZBOR REPUBLIKE SLOVENIJE - ... branjem ... XML Državni organi Javni sektor ... 27 ogledov DRŽAVNI ZBOR REPUBLIKE SLOVEN
33	evem.gov.si/evem.gov.si.371.html	točka e-VEM , Državni portal za poslovne ... na občinski ali državni cesti in čiščenje ... poteka samo po državnih oziroma po državnih in občinskih cestah ... v vlogi bančnika državnega sektorja Pogoji : ... je v izključni državni lasti s finančno ..

Results for a query: “socialno delo”
Results found in 0.32s
Raw search results found in 61.06s

Frequency	Document	Snippet
197	evem.gov.si/evem.gov.si.371.html	kmetijskem zemljišču Strojno delo s traktorjem kot ... naslednje dejavnosti/storitve : delo s traktorjem in ... ne spada : delo , ki se ... transportnih trakov za delo pod zemljo , ... in opreme za delo z vročimi kovinami ... in transporterjev za del
90	podatki.gov.si/podatki.gov.si.340.html	za svetovanje , socialni razvoj , usposabljanje ... CENTER ZA SOCIALNO DELO AJDOVŠČINA CENTER ZA SOCIALNO DELO BREŽICE CENTER ZA SOCIALNO DELO CELJE CENTER ZA SOCIALNO DELO CERKNICA CENTER ZA SOCIALNO DELO ČRNOMELJ CENTER ZA SOCIALNO DELO DOMŽALE CEN
60	evem.gov.si/evem.gov.si.29.html	oblike podjetij / Socialno podjetje (So.p ... oblike podjetij › Socialno podjetje (So.p .) Socialno podjetje (So.p .) Status socialnega podjetja lahko pridobijo ... . Glavni namen socialnega podjetništva je zaposliti ... da v okviru socialnega p
50	evem.gov.si/evem.gov.si.32.html	Mar Prispevki za socialno varnost za lastnike ... prostovoljno vključene v socialno zavarovanje , če ... Mar Prispevki za socialno varnost za samozaposlene ... Mar Prispevki za socialno varnost za družbenike ... Feb Prispevki za socialno varnost za l
32	e-uprava.gov.si/e-uprava.gov.si.31.html	za prebivanje in delo Objavljeno : 1 ...) Center za socialno delo Celje Sklep št ... Slovenije Center za socialno delo Celje Center za socialno delo Dolenjska in Bela krajina Center za socialno delo Gorenjska Center za socialno delo Južna Primorska
29	evem.gov.si/evem.gov.si.73.html	delovne pogoje , socialno varnost , ki ... je na primer delo po podjemni ali ... začasno ali občasno delo upokojencev , osebno dopolnilno delo ali kratkotrajno delo . Pridobite si ... vključen v obvezna socialna zavarovanja (pokojninsko ... in nepre

Results for a query: “social services”
Results found in 0.33s
Raw search results found in 62.11s

Frequency	Document	Snippet
15	podatki.gov.si/podatki.gov.si.407.html	kultura in šport Sociala in zaposlovanje Zdravje ... kultura in šport Sociala in zaposlovanje Zdravje ... podatkov Koristne povezave Sociala in zaposlovanje Novosti ...) Filtri Področje Sociala in zaposlovanje Občinski ... Go Državni organi Sociala
15	podatki.gov.si/podatki.gov.si.34.html	kultura in šport Sociala in zaposlovanje Zdravje ... kultura in šport Sociala in zaposlovanje Zdravje ... podatkov Koristne povezave Sociala in zaposlovanje Novosti ...) Filtri Področje Sociala in zaposlovanje Občinski ... Go Državni organi Sociala
15	podatki.gov.si/podatki.gov.si.5.html	kultura in šport Sociala in zaposlovanje Zdravje ... kultura in šport Sociala in zaposlovanje Zdravje ... podatkov Koristne povezave Sociala in zaposlovanje Novosti ...) Filtri Področje Sociala in zaposlovanje Občinski ... Go Državni organi Sociala
15	podatki.gov.si/podatki.gov.si.415.html	kultura in šport Sociala in zaposlovanje Zdravje ... kultura in šport Sociala in zaposlovanje Zdravje ... podatkov Koristne povezave Sociala in zaposlovanje Novosti ...) Filtri Področje Sociala in zaposlovanje Občinski ... Go Državni organi Sociala
15	podatki.gov.si/podatki.gov.si.153.html	kultura in šport Sociala in zaposlovanje Zdravje ... kultura in šport Sociala in zaposlovanje Zdravje ... podatkov Koristne povezave Sociala in zaposlovanje Novosti ...) Filtri Področje Sociala in zaposlovanje Občinski ... Go Državni organi Sociala
15	podatki.gov.si/podatki.gov.si.408.html	kultura in šport Sociala in zaposlovanje Zdravje ... kultura in šport Sociala in zaposlovanje Zdravje ... podatkov Koristne povezave Sociala in zaposlovanje Novosti ...) Filtri Področje Sociala in zaposlovanje Občinski ... Go Državni organi Sociala

Results for a query: “varovanje človekovih pravic”
Results found in 0.56s
Raw search results found in 62.52s

Frequency	Document	Snippet
231	evem.gov.si/evem.gov.si.371.html	javnega razpisa . Pravico udejstvovanja v lovu ... , ki si pravico udejstvovanja v lovu ... : obnova in varovanje sestojev , nega ... o rudarstvu Rudarska pravica je pravica do raziskovanja in ... razvrščа na rudarsko pravico za raziskovanje in rudar
30	e-uprava.gov.si/e-uprava.gov.si.59.html	, zakonska zveza Pravice in prejemki družine Otroški dodatek Pravice in prejemki družine ... Vloge za podaljševanje pravice do otroškega dodatka ... nadaljnji upravičenosti do pravice . Če se za omenjeno pravico zaproša prvič , ... vloga za osnovne
24	e-uprava.gov.si/e-uprava.gov.si.56.html	ne uveljavljate več pravice do starševskega dopusta ... vozniškim dovoljenjem izkazujete pravico do vožnje motornega ... za uveljavljanje svojih pravic . Potrdilo iz ... pri uveljavljanju svojih pravic priložiti tudi potrdilo ... ime je osebna pravic
21	evem.gov.si/evem.gov.si.224.html	energije , varstva človekovega zdravja in varstva ... energije , varstva človekovega zdravja in varstva ... Zakon o zasebnem varovanju Dejavnost zasebnega varovanja vključuje varovanje življenja , osebne ... varovanem območju ter varovanje premičnega
18	evem.gov.si/evem.gov.si.26.html	članske in premoženjske pravice . Osnovni kapital ... vključuje vsebino članskih pravic njenega imetnika . Članske pravice vključujejo premoženjske pravice in članske pravice v ožjem pomenu . Najpomembnejša premoženjska pravica je pravica do udeležbe
17	evem.gov.si/evem.gov.si.218.html	Izvajanje sistemov tehničnega varovanja / Dejavnosti / ... Izvajanje sistemov tehničnega varovanja Izvajanje sistemov tehničnega varovanja Dejavnost zasebnega varovanja vključuje varovanje življenja , osebne ... varovanem območju ter varovanje premič

Results for a query: “trgovina”
Results found in 0.29s
Raw search results found in 60.94s

Frequency	Document	Snippet
420	evem.gov.si/evem.gov.si.371.html	gl . 46.110 trgovina na debelo s ... gl . 10.890 trgovina na debelo z ... gl . 10.890 trgovina na debelo s ... gl . 46.380 trgovina na drobno s ... spremljevalni postopek v trgovini na debelo) ... spremljevalni postopek v trgovini na debelo) ... sp
122	podatki.gov.si/podatki.gov.si.340.html	A DENT , trgovina in storitve , ... ADRIA INVESTICIJE trgovina , posredništvo , ... za proizvodnjo , trgovino in storitve d.o.o . AHATSERVIS trgovina in storitve , ... d.o.o . ALBA trgovina in proizvodnja , ... , storitve in trgovina d.o.o . ALMA .
96	evem.gov.si/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 Nespecializira
93	evem.gov.si/evem.gov.si.21.html	eVEM › Področja Trgovina Tu boste našli ... storitev na področju trgovine . Preglejte seznam ... Seznam dejavnosti Druga trgovina na drobno v nespecializiranih prodajalnah Druga trgovina na drobno zunaj ... 47.990) Nespecializirana trgovina na debel
16	evem.gov.si/evem.gov.si.623.html	Trgovina na debelo z ... › Dejavnosti › Trgovina na debelo z ... izdelki široke porabe Trgovina na debelo z ... Sem spada : trgovina na debelo z ... izdelki ipd . trgovina na debelo s ... in deli zanja trgovina na debelo s ... knjigami , časopisi trg
15	evem.gov.si/evem.gov.si.630.html	Trgovina na drobno v ... › Dejavnosti › Trgovina na drobno v ... predmeti za gospodinjstvo Trgovina na drobno v ... spada : specializirana trgovina na drobno s pohištvom specializirana trgovina na drobno s ... za razsvetljavo specializirana trgovina