

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Aljaž Srša 63120233 in Gregor Sušnik 63120102

Obdelava, indeksiranje in poizvedovanje vsebine

Poročilo tretje seminarske naloge pri predmetu Iskanje in ekstrakcija podatkov s spleta
Asistent: asist. prof. dr. Slavko Žitnik

Povzetek

V seminarski nalogi implementirava zajemanje vsebine – besedila iz vnaprej podanih dokumentov – spletnih strani. Zajeto vsebino obdelava in indeksirava v podatkovno strukturo. Tako pripravljena zbirka podatkov služi za poizvedovanje po vsebini dokumentov. V nadaljevanju implementirava poizvedovanje in na koncu preizkusiva nekaj besednih poizvedb.

Uvod

Vsakdo se prej ali slej sreča s spletnim iskalnikom. Iskalniki so z uporabniškega vidika izjemno hitri. Ko vnesemo iskano geslo in pošljemo zahtevo, nam v nekaj desetinkah sekunde vrne seznam zadetkov. Le-ta pa običajno vsebuje nekaj sto milijonov zadetkov. Nadvse osupljivo je, kako hitro predvsem pa učinkovito je delovanje iskalnikov.

V tretji seminarski nalogi smo se lotili implementacije indeksiranja vsebine spletnih strani in implementacije iskalnika. Ideja je sprogramirati algoritem, ki bo v dokumentu izluščil vsebino in jo ustrezno obdelal ter na koncu shranil v podatkovno zbirko. V drugem delu seminarske pa je ideja implementirati algoritem – iskalnik, ki na podlagi vpisane besedne zveze le to najprej ustrezno obdeli in vrne rezultate iz podatkovne zbirke nazaj uporabniku. Pri tem mora iskalnik vrniti čim bolj relevantne zadetke.

Obdelava vsebine in indeksiranje dokumentov

Algoritem na začetku sestavi seznam. V seznamu se nahajajo relativne poti vseh dokumentov, ki jih bo obdelal. Nato za vsak path dokumenta v seznamu s pomočjo funkcije *beautifulsoup* olepša. Za procesiranje vzameva le besedilo, ki se nahaja znotraj značke *body*. Temu najprej odstraniva vse skripte in ga nato tokenizirava. Vse besede pretvoriva v majhne črke in filtrirava s pomočjo seznama t.i. *stopword*, omenjeni seznam je del paketa *nlTK*. S pomočjo dodatnega seznama, ki predstavlja znake ločil, oklepajev in podobno, dokončno filtrirava besede. Tako obdelana vsebina dokumenta je nared za nadaljnji proces. S pomočjo funkcije *FreqDist()* pridobiva frekvenco pojavitve za vsako besedo v obdelani vsebini dokumenta. Z uporabo regularnega izraza »r"\W%s\W" % word« in uporabo funkcije *re.finditer()* dobiva seznam indeksov. Indeks pove, na katerem mestu v dokumentu se nahaja beseda.

Vse informacije shraniva v podatkovno zbirko. Ta sestoji iz dveh tabel. *IndexWord*, ki se

obnaša kot slovar. V njej so shranjene vse besede celotnega prečiščenega korpusa dokumentov. Besede se shranijo samo enkrat. Kasnejše pojavitve iste besede, se ne shranijo. V tabeli *Posting* pa se shrani beseda, dokument, v katerem se nahajajo besede, število pojavitev besede v dokumentu in seznam indeksov. Kot že opisano, seznam predstavlja številke, ki povedo na katerem mestu v dokumentu se nahaja beseda. Ko algoritem obdela vse dokumente, se prvi del seminarske naloge zaključi.

Podatkovna zbirka

V tabeli *indexWord* je shranjenih 47.529 besed, v tabeli *Posting* pa je 393.609 zapisov. Največjo frekvenco ima beseda *proizvodnja*, ki se v dokumentu *evem.gov.si.371.html* pojavi kar 2.266 krat. Sledita ji besedi *spada*, ki se v enem dokumentu pojavi 1.338 krat in beseda *dejavnosti*, ki se v enem dokumentu pojavi 1.287 krat.

V celotnem korpusu se največkrat pojavi beseda *podatkov*, in sicer 11.045 krat. Sledita ji besedi *slovenije*, ki se pojavi 9.926 krat in *republike*, ki se pojavi 8.570 krat.

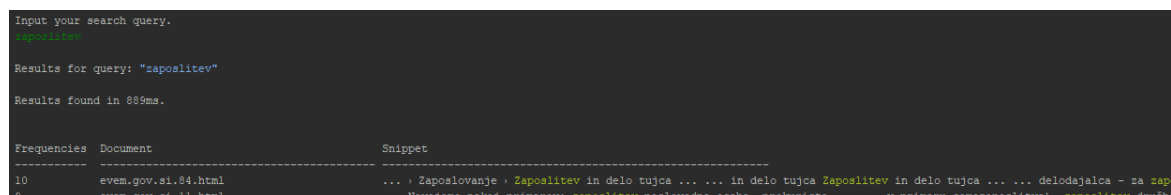
Iskanje vsebine

Iskanje vsebine sva implementirala na dva načina. V prvem načinu se poizvedovanje dokumentov izvaja s pomočjo podatkovne zbirke. V drugem načinu pa se iskanje izvaja direktno v dokumentih in se ne uporablja podatkovne zbirke.

Iskanje vsebine s pomočjo podatkovne zbirke

Algoritem vneseno besedo oz. besedno zvezo razbije na seznam besed. Nato se izvede poizvedba SQL. Ta sešteje število pojavitev posameznih vnesenih besed – frekvence v posameznem dokumentu, in združi seznime indeksov za posamezen dokument. Uredi se tudi vrstni red podatkov, in sicer se podatki razvrstijo v padajočem vrstnem redu po frekvenci. Tako pridobljeni podatki iz podatkovne zbirke se izpišejo uporabniku kot rezultat. Pri izpisu se izpiše frekvenca pojavitve besede, dokument in odrezek. Tega sva pri izpisu omejila tako, da se prikazuje največ 5 odrezkov na dokument. Izračun odrezka se naredi tako, da iz vsebine dokumenta, ki ga dobiva iz značke *body* (ko je že odstranjena navlaka) poišče mesto (indeks), kjer se nahaja iskana beseda. Nato od trenutnega indeksa vzameva tri elemente naprej in nazaj ter vse skupaj sestaviva skupaj s tremi pikami, če je to potrebno. Upošteva tudi primer, ko je iskana beseda na začetku ali koncu. Spodnja Slika 1 prikazuje primer

izpisa za iskano besedo 'zaposlitev'.



```
Input your search query.
zaposlitev

Results for query: "zaposlitev"

Results found in 889ms.

Frequencies Document Snippet
-----
10 evem.gov.si.84.html ... . Zaposlovanje . Zaposlitev in delo tujca ... . in delo tujca Zaposlitev in delo tujca ... . delodajalca - za zap
9 evem.gov.si.11.html ... . Navajamo nekaj primerov: zaposlitev poslovodne osebe, prokurista ... . v primeru samozaposlitve), zaposlitev družb
```

Slika : Demonstracija iskalnik za vnešeno geslo 'zaposlitev'

Iskanje vsebine v dokumentih

Algoritem tudi tukaj vneseno besedo oz. besedno zvezo razbije v seznam besed. Nato začne izvajati iskanje. To počne tako, da vsak dokument najprej prečisti nato pa pregleda, če je iskana beseda (ali več besed) vsebovana v dokumentu. Seznam, v katerega se shranjujejo

rezultati – seštevek frekvence za posamezen dokument, naziv dokumenta in seznam indeksov se tekom obdelave dokumentov polni s podatki in po kar nekaj pretečenega časa se prikažejo rezultati.

Rezultati

S pomočjo implementiranega iskalnika sva preizkusila delovanje z uporabo različnih besed. Za spodnje iskanje sva dobila naslednje rezultate:

Results for query: "predelovalne dejavnosti"

Results found in 84ms.

Frequencies	Document	Snippet
1291	evem.gov.si.371.html	... iskanje ustrezne šifre dejavnosti /storitve in informacij pogojih za op
75	evem.gov.si.377.html	... Defektolog v zdravstveni dejavnosti Dekan oziroma direktor Dietetik v z
40	podatki.gov.si.340.html	... - NOSILEC DOPOLNILNE DEJAVNOSTI NA KMETIJI BREGAR Šport CENTER INTERESN
38	evem.gov.si.452.html	... e-VEV eVEM » Dejavnosti » Druge storitvene » Druge storitvene dejavnosti
31	evem.gov.si.653.html	... Dovoljenje za opravljanje dejavnosti specializirane prodajalne z radijs
30	evem.gov.si.398.html	... usmerjene na opravljanje dejavnosti (npr.: pripravljalna dela, za namen
28	evem.gov.si.72.html	... od dohodka iz dejavnosti Davek od dohodka od dohodka iz dejavnosti Ko z
23	evem.gov.si.442.html	... e-VEV eVEM » Dejavnosti » Dejavnosti za » Dejavnosti » Dejavnosti za nego
18	evem.gov.si.28.html	... za opravljanje gospodarske dejavnosti . Lastnosti zasebnega zavoda niso
17	evem.gov.si.265.html	... e-VEV eVEM » Dejavnosti » Proizvodnja mesa, SKD šifra zajema dejavnosti
17	evem.gov.si.460.html	... e-VEV eVEM » Dejavnosti » Druge nerazvrščene Druge nerazvrščene predel
16	evem.gov.si.266.html	... e-VEV eVEM » Dejavnosti » Proizvodnja mesnih SKD šifra zajema dejavnosti
16	evem.gov.si.272.html	... e-VEV eVEM » Dejavnosti » Storitve za SKD šifra zajema dejavnosti in sto
16	evem.gov.si.276.html	... e-VEV eVEM » Dejavnosti » Storitve za SKD šifra zajema dejavnosti in sto
16	evem.gov.si.392.html	... redno poslovanje. 6. Dejavnosti podjetja Ob ustanovitvi prihodkov) in d

- Vpisano geslo: predelovalne dejavnosti

Results for query: "trgovina"

Results found in 702ms.

Frequencies	Document	Snippet
364	evem.gov.si.371.html	... organizacij, gl. 46.110 trgovina na debelo s juh, gl. 10.890 trgovina na debelo z
96	evem.gov.si.651.html	... Druga govedoreja Druga trgovina na drobno v specializiranih prodajalnah Druga trgovina
92	evem.gov.si.21.html	... eVEM » Področja Trgovina Tu boste našli Seznam dejavnosti Druga trgovina na drobno v ..
82	podatki.gov.si.340.html	... d.o.o. A DENT, trgovina in storitve, d.o.o. d.o.o. ADRIA INVESTICIJE trgovina , posredni
13	evem.gov.si.623.html	... » Dejavnosti » Trgovina na debelo z izdelki široke porabe Trgovina na debelo z p
12	evem.gov.si.329.html	... » Dejavnosti » Trgovina na debelo z in sanitarno opremo Trgovina na debelo z opr
12	evem.gov.si.630.html	... » Dejavnosti » Trgovina na drobno v predmeti za gospodinjstvo Trgovina na drobno v
10	evem.gov.si.320.html	... » Dejavnosti » Trgovina na debelo s napravami za ogrevanje Trgovina na debelo s
10	evem.gov.si.327.html	... » Dejavnosti » Trgovina na debelo z napravami in opremo Trgovina na debelo z opr
10	evem.gov.si.622.html	... » Dejavnosti » Trgovina na debelo z električnimi gospodinjstskimi napravami Trgovina na de
9	evem.gov.si.328.html	... » Dejavnosti » Trgovina na debelo s in plinastimi gorivi Trgovina na debelo s ta
9	evem.gov.si.343.html	... » Dejavnosti » Trgovina na drobno v prodajalnah s tekstilom Trgovina na drobno v
9	evem.gov.si.620.html	... » Dejavnosti » Trgovina na debelo z napravami in deli Trgovina na debelo z deli
8	evem.gov.si.316.html	... » Dejavnosti » Trgovina na debelo s in rudami (46.720) Trgovina na debelo s (46.
8	evem.gov.si.323.html	... » Dejavnosti » Trgovina na debelo s debelo s tekstilom Trgovina na debelo s teks
8	evem.gov.si.334.html	... » Dejavnosti » Trgovina na debelo z semeni in krmo Trgovina na debelo z krmo Sem
7	evem.gov.si.315.html	... » Dejavnosti » Trgovina na debelo s in materiali (46.460) Trgovina na debelo s (
7	evem.gov.si.319.html	... » Dejavnosti » Trgovina na debelo s stroji, priključki, opremo Trgovina na debelo s
7	evem.gov.si.336.html	... » Dejavnosti » Trgovina na debelo s stroji, priključki, opremo Trgovina na debelo s

- Vpisano geslo: trgovina

Results for query: "social services"

Results found in 69ms.

Frequencies	Document	Snippet
5	e-uprava.gov.si.45.html	... culture Labour, retirement Social services, health, death ...
5	e-uprava.gov.si.9.html	... culture Labour, retirement Social services, health, death ...
1	evem.gov.si.661.html	... Records and Related Services (AJ PES) and the ...
1	podatki.gov.si.340.html	... recreation and spa services ltd. TERME MARIBOR, ...

- Vpisano geslo: social services

Results for query: "zaposlitev"

Results found in 53ms.

Frequencies	Document	Snippet
10	evem.gov.si.84.html	... » Zaposlovanje » Zaposlitev in delo tujca ... in delo tujca Zaposlitev in delo tujca ...
9	evem.gov.si.111.html	... Navajamo nekaj primerov: zaposlitev poslovodne osebe, prokurista ... v prime
7	evem.gov.si.371.html	... imenu igralcev iščejo zaposlitev pri filmskih, gledaliških ... opravljanje o
6	podatki.gov.si.105.html	... csv Dovoljeno število zaposlitev in zaposlenih v ... csv Dovoljeno število d
3	evem.gov.si.29.html	... del, subvencije za zaposlitev, nepovratna sredstva za ... zaposlijo ali ohra
2	e-uprava.gov.si.56.html	... ki mu omogoča zaposlitev na določenem poklicnem ... in delodajalcem. Prva za
2	evem.gov.si.35.html	... Pridobitev dovoljenja za zaposlitev tujca na podlagi ... Pridobitev dovoljen
2	evem.gov.si.405.html	... za drugo ustrezno zaposlitev na podlagi nove ... ponujeno drugo ustrezno zap
2	evem.gov.si.6.html	... Pridobitev dovoljenja za zaposlitev tujca na podlagi ... Pridobitev dovoljen
1	e-uprava.gov.si.19.html	... zaščita Pogoji za zaposlitev v Slovenski vojski, ...
1	evem.gov.si.366.html	... najpomembnejšo možnost za zaposlitev, podjetjem pa nudijo ...
1	evem.gov.si.382.html	... oz. uživanje pravice (zaposlitev s polnim ali ...
1	evem.gov.si.398.html	... izdajo dovoljenja za zaposlitev tujca, podatke o ...
1	evem.gov.si.651.html	... Vodenje, varstvo in zaposlitev pod posebnimi pogoji ...
1	evem.gov.si.73.html	... delo v tujino Zaposlitev in delo tujca ...
1	podatki.gov.si.340.html	... za rehabilitacijo in zaposlitev invalidov Maribor d.o.o. ...

- Vpisano geslo: zaposlitev

Results for query: "kazenski zakonik"

Results found in 62ms.

Frequencies	Document	Snippet
5	e-uprava.gov.si.25.html	... ni zabeležena v kazenski evidenci. Potrdilo izkazuje ... neobsojeno. Pravna podlaga Ka
4	podatki.gov.si.77.html	... katere je bil kazenski postopek pri državnem ... katere je bil kazenski postopek pri d
4	podatki.gov.si.85.html	... katere je bil kazenski postopek pri državnem ... katere je bil kazenski postopek pri d
4	podatki.gov.si.86.html	... katere je bil kazenski postopek pri državnem ... katere je bil kazenski postopek pri d
4	podatki.gov.si.87.html	... katere je bil kazenski postopek pred senatom ... katere je bil kazenski postopek pred
4	podatki.gov.si.88.html	... katere je bil kazenski postopek pred senatom ... katere je bil kazenski postopek pri d
3	podatki.gov.si.336.html	... katere je bil kazenski postopek pri državnem ... katere je bil kazenski postopek pri d
3	podatki.gov.si.358.html	... katere je bil kazenski postopek pri državnem ... katere je bil kazenski postopek pri d
3	podatki.gov.si.364.html	... dejanju in glavni kazenski sankciji (tudi pogojno ... odgovorne po glavni kazenski san
3	podatki.gov.si.73.html	... katere je bil kazenski postopek pri državnem ... katere je bil kazenski postopek pri d
3	podatki.gov.si.75.html	... katere je bil kazenski postopek pri državnem ... katere je bil kazenski postopek pri d
2	evem.gov.si.26.html	... družbah , Obligacijski zakonik , Zakon o ... gospodarskih družbah Obligacijski zakonik
2	evem.gov.si.371.html	... zakonom, ki ureja kazenski postopek. Kandidat za ... Pogoji: - Pomorski zakonik Uprava
1	evem.gov.si.227.html	... zakonom, ki ureja kazenski postopek. Kandidat za ...
1	evem.gov.si.239.html	... Pravne podlage Pomorski zakonik (PZ) Pravilnik o ...
1	evem.gov.si.240.html	... Pravne podlage Pomorski zakonik (PZ) Uredba o ...
1	evem.gov.si.241.html	... Pravne podlage Pomorski zakonik (PZ) Pravilnik o ...
1	evem.gov.si.242.html	... Pravne podlage Pomorski zakonik (PZ) Pravilnik o ...
1	evem.gov.si.244.html	... Pravne podlage Pomorski zakonik (PZ) Uredba o ...
1	evem.gov.si.533.html	... Pravne podlage Pomorski zakonik (PZ) Pravilnik o ...
1	evem.gov.si.534.html	... Pravne podlage Pomorski zakonik (PZ) Pravilnik o ...
1	evem.gov.si.535.html	... Pravne podlage Pomorski zakonik (PZ) Uredba o ...
1	evem.gov.si.536.html	... Pravne podlage Pomorski zakonik (PZ) Uredba o ...
1	evem.gov.si.541.html	... ostankov tovara Pomorski zakonik (PZ) Zakon o ...
1	evem.gov.si.75.html	... Pravna podlaga Obligacijski zakonik (člen. 619-648) Pomoč ...
1	podatki.gov.si.404.html	... katere je bil kazenski postopek pri državnem ...

- Vpisano geslo: kazenski zakonik

Results for query: "Ministrstvo za javno upravo"

Results found in 115ms.

Frequencies	Document	Snippet
615	evem.gov.si.371.html	... kmetijskih rastlin Vir: Ministrstvo za kmetijstvo, gozdarstvo ... oziroma živil Vir:
139	podatki.gov.si.340.html	... z odpadki, d.o.o., javno podjetje CESTNO PODJETJE ... ORDINACIJA ČISTA NARAVA, javno
55	podatki.gov.si.340.html	... (227) REPUBLIKA SLOVENIJA, MINISTRSTVO ZA DELO, DRUŽINO, ... ENAKE MOŽNOSTI (61) MIN
53	podatki.gov.si.31.html	... NACIONALNI INŠTITUT ZA JAVNO ZDRAVJE (411) BANKA ... (58) MINISTRSTVO ZA JAVNO UPRAVO
45	podatki.gov.si.69.html	... Organizacije Podrobnosti Organizacija: MINISTRSTVO ZA JAVNO UPRAVO ... ZA JAVNO UPRAVO
42	e-uprava.gov.si.55.html	... svetu. Vlogo ureja: Ministrstvo za javno upravo ... anonimno. Vlogo ureja: Ministrst
37	podatki.gov.si.368.html	... subjekti (1) Organizacija MINISTRSTVO ZA NOTRANJE ZADEVE ... NOTRANJE ZADEVE (52) MIN
36	podatki.gov.si.51.html	... organi (11) Organizacija MINISTRSTVO ZA JAVNO UPRAVO ... mesecih 1797 ogledov MINIST
34	podatki.gov.si.2.html	... NACIONALNI INŠTITUT ZA JAVNO ZDRAVJE (407) BANKA ... (12) MINISTRSTVO ZA JAVNO UPRAVO
34	podatki.gov.si.335.html	... NACIONALNI INŠTITUT ZA JAVNO ZDRAVJE (407) BANKA ... (12) MINISTRSTVO ZA JAVNO UPRAVO
34	podatki.gov.si.48.html	... REPUBLIKE SLOVENIJE (198) MINISTRSTVO ZA JAVNO UPRAVO ... JAVNO UPRAVO (11) MINISTR
33	podatki.gov.si.63.html	... NACIONALNI INŠTITUT ZA JAVNO ZDRAVJE (407) BANKA ... (12) MINISTRSTVO ZA JAVNO UPRAVO
33	podatki.gov.si.97.html	... organi (9) Organizacija MINISTRSTVO ZA JAVNO UPRAVO ... mesecih 1797 ogledov MINIST
31	podatki.gov.si.43.html	... organi (22) Organizacija MINISTRSTVO ZA JAVNO UPRAVO ... JAVNO UPRAVO (9) MINISTRSTVO
27	podatki.gov.si.90.html	... organi (10) Organizacija MINISTRSTVO ZA JAVNO UPRAVO ... mesecih 1797 ogledov MINIST
27	podatki.gov.si.91.html	... organi (10) Organizacija MINISTRSTVO ZA JAVNO UPRAVO ... mesecih 1797 ogledov MINIST
26	podatki.gov.si.424.html	... NACIONALNI INŠTITUT ZA JAVNO ZDRAVJE NACIONALNI INŠTITUT ... NACIONALNI INŠTITUT ZA
24	podatki.gov.si.93.html	... organi (8) Organizacija MINISTRSTVO ZA JAVNO UPRAVO ... JAVNO UPRAVO (6) MINISTRSTVO
23	evem.gov.si.653.html	... na obnovljive vire Javno pooblastilo Javno pooblastilo ... vire Javno pooblastilo Jav
23	podatki.gov.si.41.html	... subjekti (1) Organizacija MINISTRSTVO ZA JAVNO UPRAVO ... JAVNO UPRAVO (6) MINISTRST
22	podatki.gov.si.168.html	... NACIONALNI INŠTITUT ZA JAVNO ZDRAVJE (407) VLADA ... NACIONALNI INŠTITUT ZA JAVNO ZD
22	podatki.gov.si.38.html	... organi (18) Organizacija MINISTRSTVO ZA JAVNO UPRAVO ... REPUBLIKE SLOVENIJE (2) MIN
22	podatki.gov.si.428.html	... NACIONALNI INŠTITUT ZA JAVNO ZDRAVJE (407) VLADA ... NACIONALNI INŠTITUT ZA JAVNO ZD

- Vpisano geslo: Ministrstvo za javno upravo

Pri uporabi iskanja s pomočjo podatkovne zbirke so bili rezultati vidni takoj. Ko pa sva enake poizvedbe poskušala pridobiti na način direktnega iskanja v dokumentih pa sva za rezultate morala čakati ogromno časa. Razlika med iskalnima načinoma je ogromna.