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 v naprej 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

Vsak kdo 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 obdela 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 števke, ki povedo na katerem mestu v dokumentu se nahaja beseda. Ko algoritem obdela vse dokumente, se prvi del seminarske zaključi.

Podatkovna zbirka

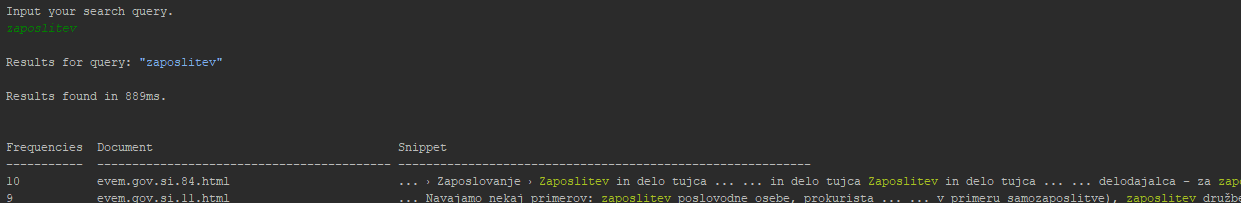
V tabeli *indexWord* je shranjenih 47.529 besed, v tabeli *Posting* pa je 393.609 zapisov. Največjo frekvenco ima beseda *proizvodnja*, ki se v dokumentu *evem.gov.si.371.*html pojavi kar 2.266 krat. Sledita ji besedi *spada*, ki se v enem dokumentu pojavi 1.338 krat in beseda *dejavnosti*, ki se v enem 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 vsebino 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 sezname indeksov za posamezen dokument. Uredi se tudi vrstni red podatkov, in sicer se podatki razvrstijo v padajočem vrstnem redu. 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) poiščeva 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. Upoštevava tudi primer, ko je iskana beseda na začetku ali koncu. Spodnja Slika 1 prikazuje primer izpisa za iskano besedo 'zaposlitev'.  
  


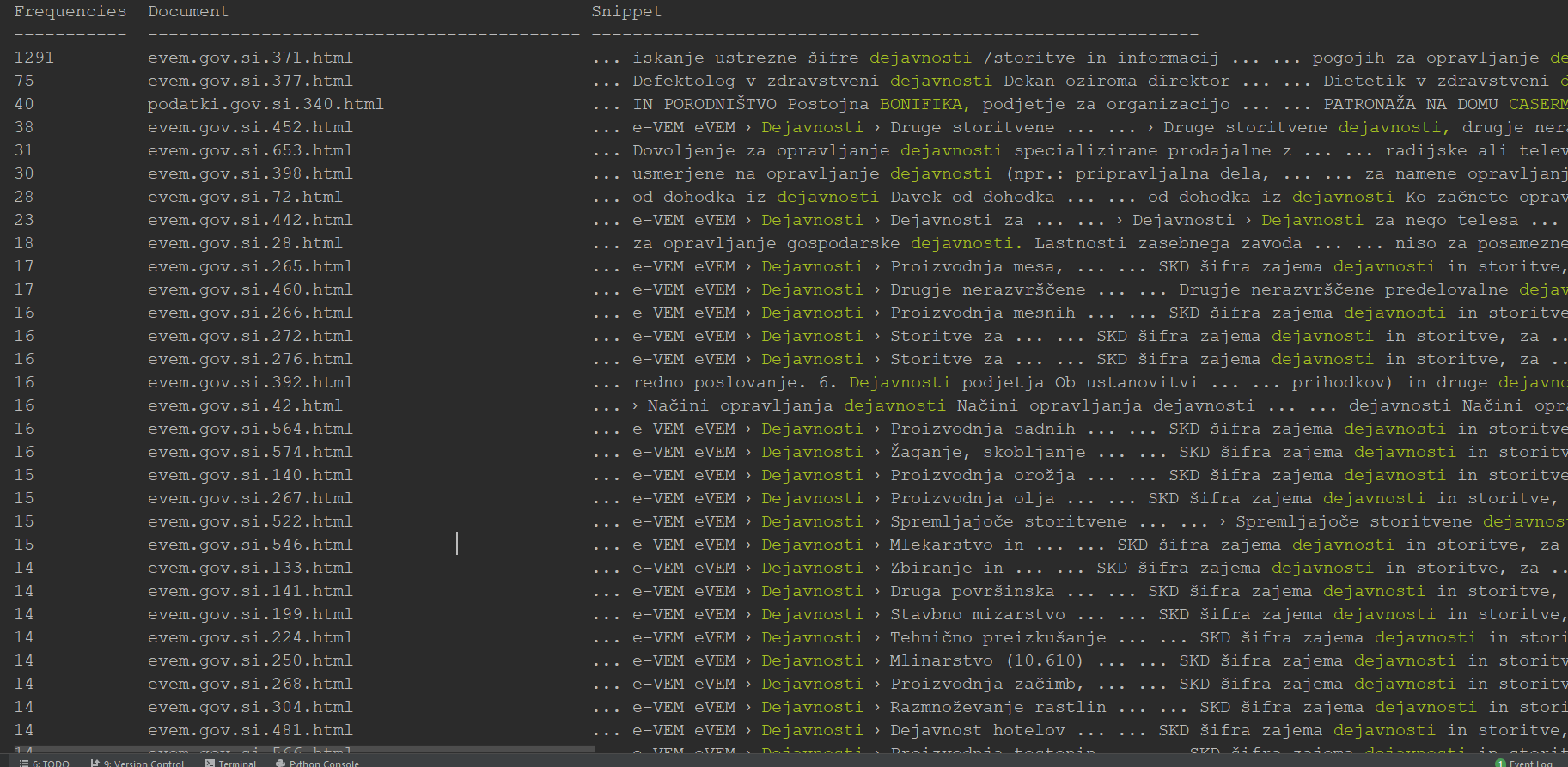
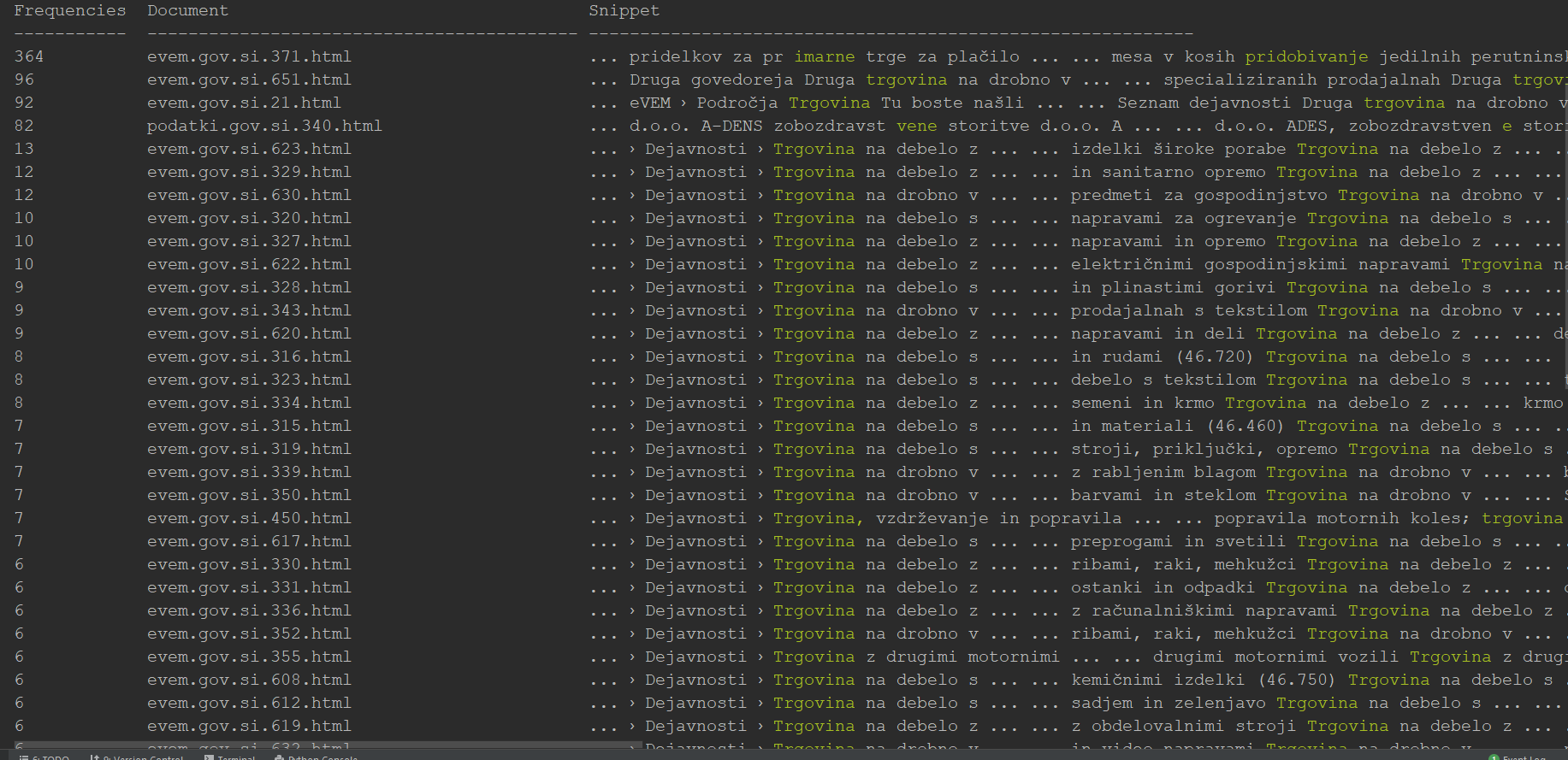
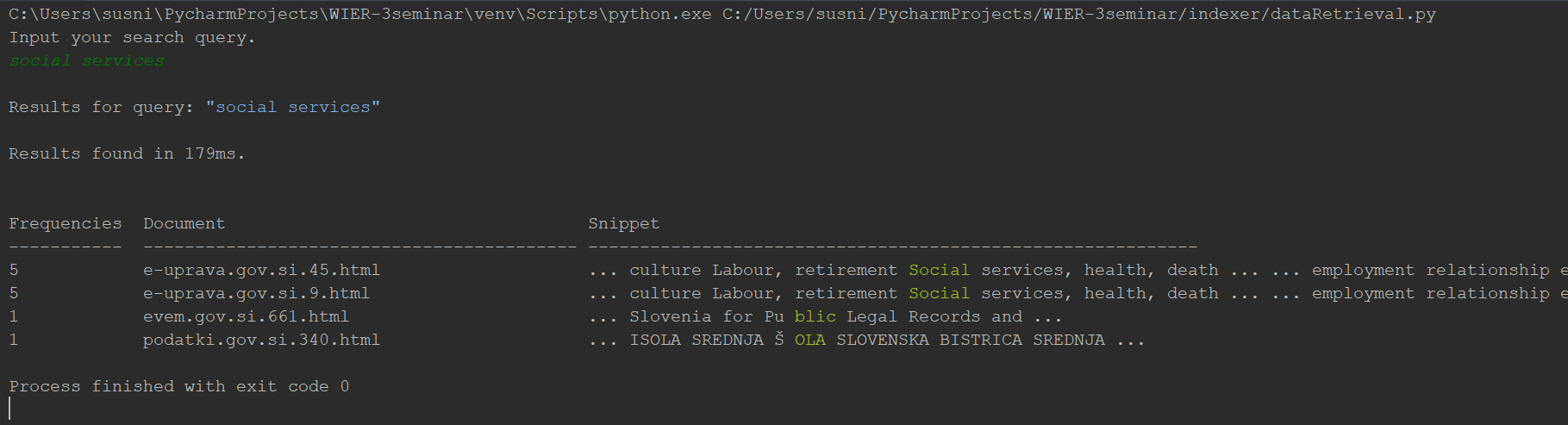
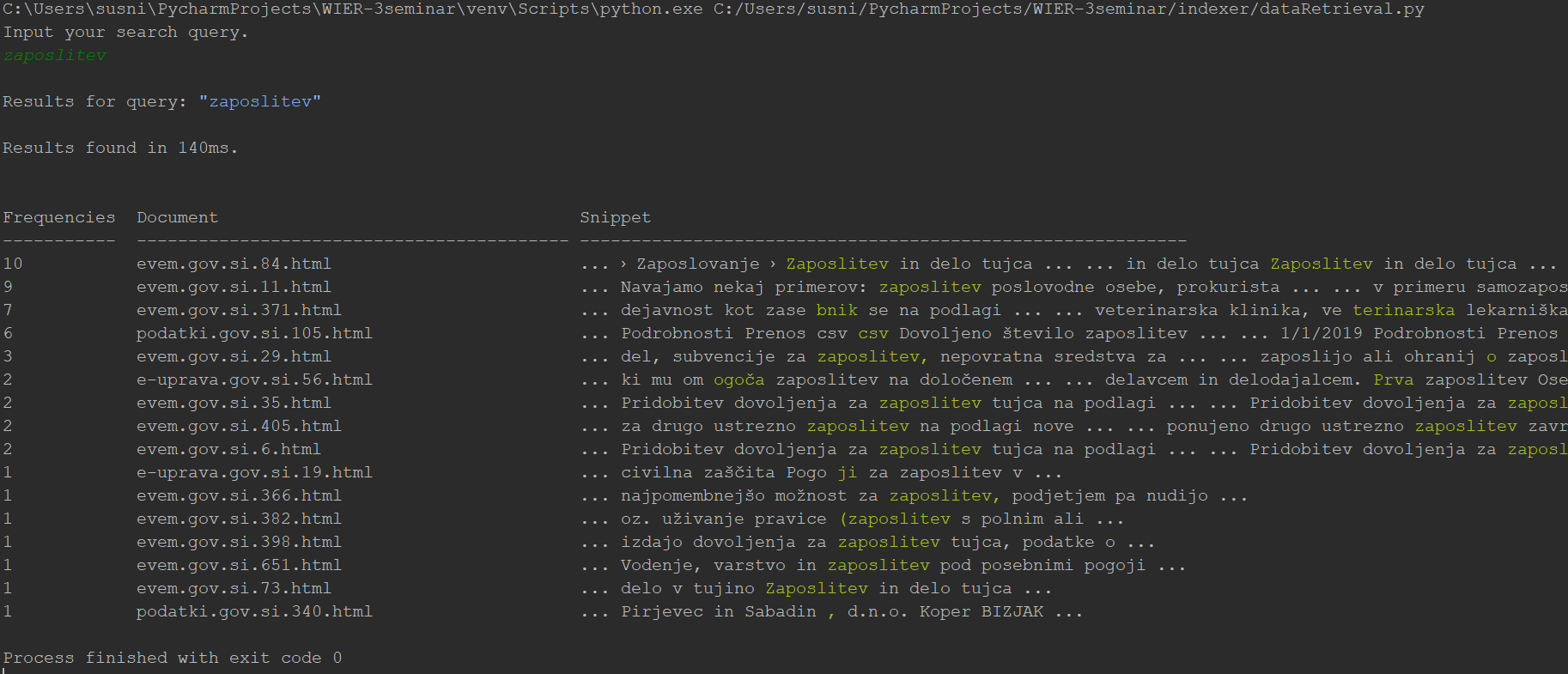
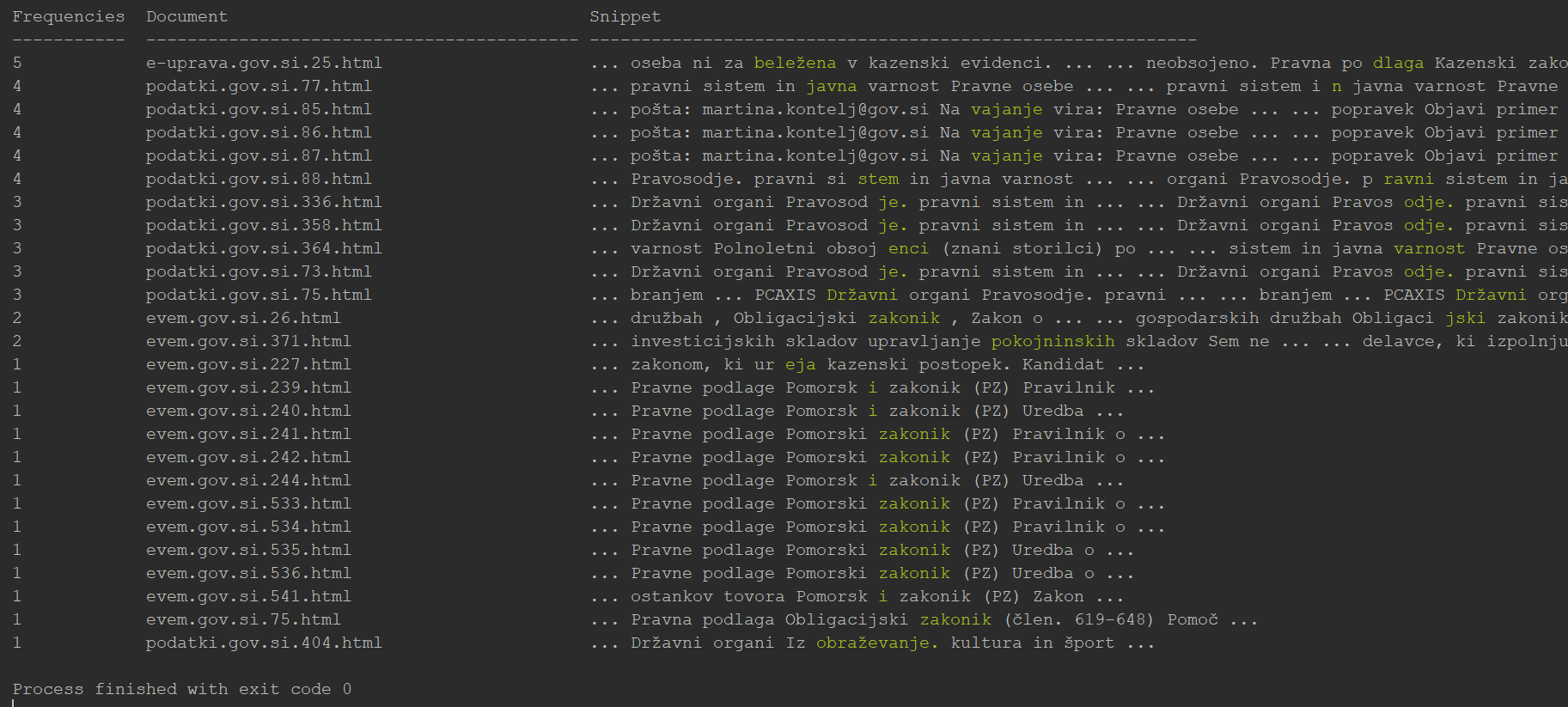
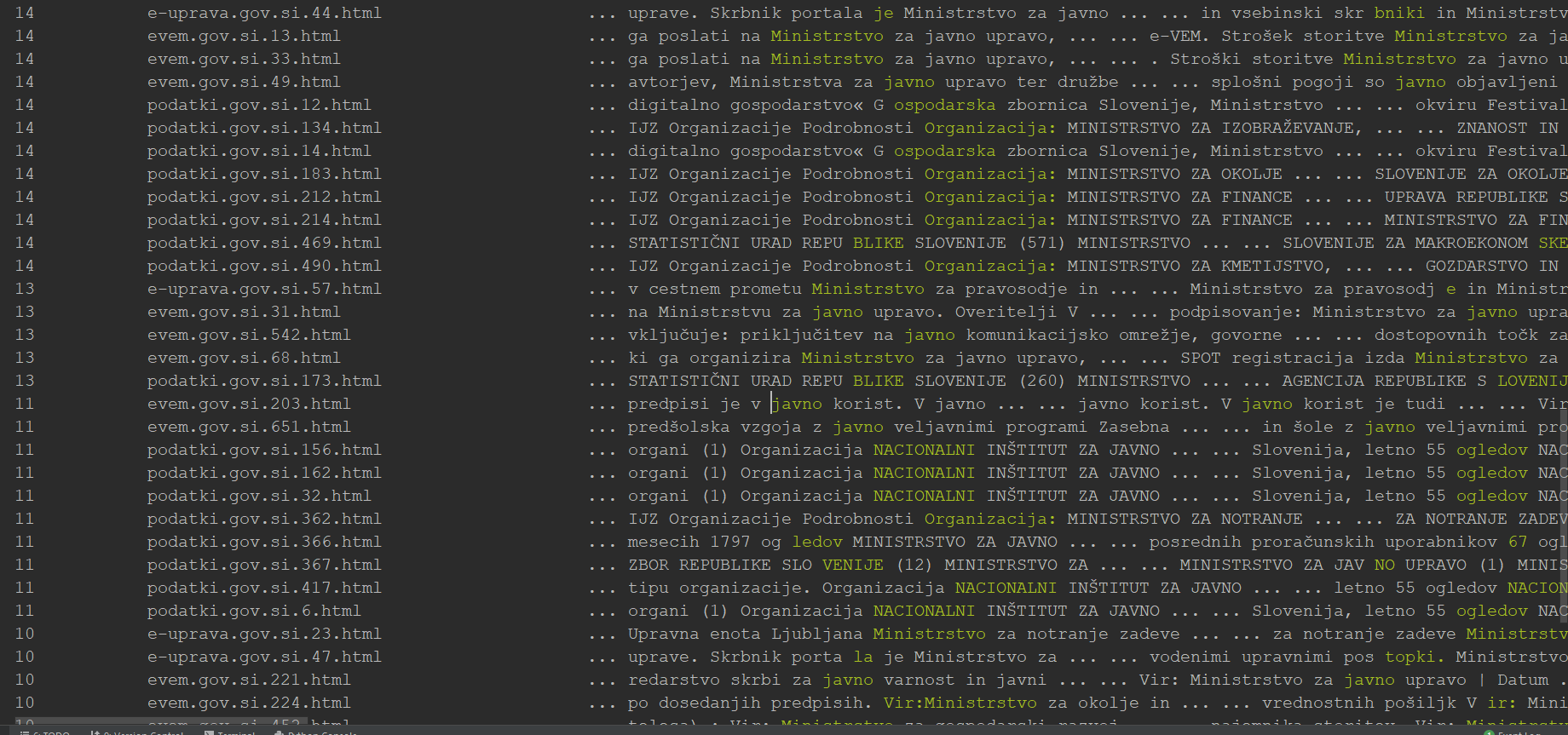
Slika : Demonstracija iskalnik za vnešeno geslo 'zaposlitev'

**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:

* Vpisano geslo: predelovalne dejavnosti
* Vpisano geslo: trgovina
* Vpisano geslo: social services
* Vpisano geslo: zaposlitev
* Vpisano geslo: kazenski zakonik
* Vpisano geslo: Ministrstvo za javno upravo

Pri uporabi iskanja s pomočjo podatkovne zbirke so rezultati bili 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.