1 Implementacija

1.1 Razčlenjevanje dokumenta

Dokument se razčleni na posamezne člene (to so besede, številke, znake, ipd.) s funkcijo `nltk.word_tokenize` z uporabo slovenskega jezika. Posamezne člene še dodatno obdelamo tako, da zavržemo vse člene, ki nimajo nobene črke, dodatno pa odstranimo še neželene znake na začetku ali koncu besede, kot so navednice, pomišljaji, podčrtaji, ipd. (na takšen način iz člena *"_klepar"* dobimo končen rezultat *klepar*, člen *332.55* pa bo zavržen, saj ne vsebuje nobene črke). V kolikor se izkaže, da je prečiščen člen na seznamu nepomembnih besed (angl. *stop-word*), se ga ravno tako zavrže. V kolikor člen po koncu obdelave ni zavržen, se ga shrani v podatkovno bazo, kot opisano v podpoglavju 1.2, oziroma se ga uporabi pri iskalnem algoritmu brez invertnega indeksa, kot opisano v podpoglavju 1.4.

1.2 Indeksiranje

Najprej smo naredili novo bazo podatkov z dvema tabelama IndexWord in Posting. Za tem smo obdelali vsako HTML datoteko posebej. Postopek razčlenjevanja besed iz besedila HTML datoteke je opisan v podpoglavju 1.1. Ko smo pridobili besede smo le te vnesli v tabelo IndexWord. Prešteli smo še vse pojavitve besede, si zapisali indexe na katerih mestih se ta beseda v besedilu nahaja ter vpisali pridobljene podatke v tabelo Posting.

1.3 Iskanje z invertnim indeksom

Za iskanje pojavitev poizvedbe moramo imeti najprej napolnjeno bazo, kar je bilo opisano v prejšnjem poglavju. Skripta za iskanje v podatkovni bazi sprejme en argument, ki pa je lahko sestavljen iz več besed. Ko uporabnik zahteva iskanje vzamemo besede in iz njih ustvarimo poizvedno. Poizvedba išče pojavitve besed v stolpcu *word*, sešteva stolpec *frequency* in skupaj združi v en stolpec informacijo na katerih mestih v dokumentu se ta beseda pojavi. Ta informacija (mesta kjer je beseda) je uporabljena kasneje pri prikazu rezultatov. Poizvedba je združena (*group by*) po imenih dokumentov. Torej na koncu dobimo rezultat kjer je za vsak dokument kej se beseda pojavi vsota vseh pojavitev ter mesta kjer se ta beseda pojavi v originalnem dokumentu (datoteki). Rezultati so urejeni po vsoti frekvence pojavitev v padajočem vrstnem redu. Za prikaz osnutkov iz originalnih dokumentov, najprej odpremo dokument se pomaknemo na mesto kjer vemo da se iskana beseda nahaja in pridobimo določeno število besed pred in po njej (natančneje opisano v podpoglavju 1.5). Te nato izpišemo na standardni izhod.

1.4 Iskanje brez invertnega indeksa

Skripta najprej pridobi seznam vseh HTML datotek, ki se nahajajo v mapi `pa/pages`, nato zaporedoma vsako datoteko posebej razčleni, kot opisano v podpoglavju 1.1. Med razčlenjevanjem se shranjuje seznam indeksov besed, ki se nahajajo v poizvedbi uporabnika (tj. če se razčlenjena beseda iz dokumenta ujema z vsaj eno besedo iz poizvedbe, se shrani njen razčlenjen indeks iz dokumenta). V kolikor seznam indeksov ni prazen ob koncu obdelave dokumenta, se izdela izsek rezultatov, kot opisano v podpoglavju 1.5. Med izvajanjem skripte se hrani dokumente v katerih smo našli vsaj 1 zadetek skupaj s številom zadetkov ter pripadajočimi izseki. Ko so preiskani vsi dokumenti, se ustavi merjenje časa izvajanja skripte ter izpiše najdene rezultate.

1.5 Izdelava izseka rezultatov

Med postopkom iskanja gradimo seznam indeksov členov, ki se ujemajo s katerokoli izmed besed poizvedbe, z njegovo pomočjo ter z uporabo seznama členov pa gradimo izseke rezultatov. V osnovi algoritem sestavi izsek tako, da izpiše 3 člene pred in po najdeni besedi, vendar smo opazili, da se pri takšnem delovanju določeni deli besedila podvajajo, npr. pri iskanju besede "sistem" v besedilu "Državni portal in sistem SPOT je edinstven in unikaten sistem za poslovne subjekte in samostojne podjetnike" bi prejeli izseke "... Državni portal in sistem SPOT je edinstven in ... edinstven in unikaten sistem za poslovne subjekte ...".
Zaradi tega smo dodatno omejili gradnjo izsekov tako, da če pride do prekrivanja določenih delov, se le-ti združijo v en sam izsek. Zato so naši izseki občasno daljši, izpis iz prejšnjega primera pa bi izgledal tako:
"... Državni portal in sistem SPOT je edinstven in unikaten sistem za poslovne subjekte ...".

2 Podatkovna baza

V podatkovni bazi v tabeli Indexwords imamo 35290 besed.
V spodnji levi tabeli so prikazani dokumenti z največ ponovitvami iste besede. V desni pa besede, ki so uporabljene v največ dokumentih.

Dokument	Beseda	Št. besed
evem.gov.si.371.html	proizvodnja	2266
evem.gov.si.371.html	spada	1338
evem.gov.si.371.html	dejavnosti	1287
podatki.gov.si.340.html	skupnost	809
podatki.gov.si.340.html	krajevna	754

Beseda	Št. dokumentov
uporabe	1399
pogoji	1398
domov	1384
portalu	1367
slovenije	1363

V spodni tabeli so prikazane najbolj pogoste besede v vseh dokumentih skupaj.

Beseda	Št. dokumentov	Št. besed
podatkov	864	11090
slovenije	1363	10507
republike	1165	8583
dejavnosti	754	6094
podatki	734	5809

3 Rezultati poizvedb

Pri iskanju brez invertnega indeksa pridobimo rezultate v 36-37 sekundah, pri iskanju z indeksom pa rezultate pridobimo v manj kot 10ms, vendar se ti izpišejo po tem, ko izdelamo vse izseke rezultatov (to traja običajno 1-5 sekund). Izpisi so sledeči:

Results for query: "predelovalne dejavnosti"
754 results found in 6ms

Freq.	Document	Snippets (limited to 3)
1291	evem.gov.si.371.html	... za infrastrukturo C PREDELOVALNE DEJAVNOSTI 10 Proizvodnja ... 32 Druge raznovrstne predelovalne dejavnosti 32.110 Kovanje ... 32.990 Drugje nerazvrščene predelovalne dejavnosti Sem spada ...
75	evem.gov.si.377.html	... Defektolog v zdravstveni dejavnosti Dekan oziroma direktor ... Dietetik v zdravstveni dejavnosti Dimnikar Diplomirana medicinska ... I v zdravstveni dejavnosti Laboratorijski sodelavec II v zdravstveni dejavnosti Laboratorijski tehnik Ladijski ...
40	podatki.gov.si.340.html	... - NOSILEC DOPOLNILNE DEJAVNOSTI NA KMETIJI BREGAR ... šport CENTER INTERESNIH DEJAVNOSTI PTUJ CENTER JUDOVSKE ... ŠOLSKIH IN OBŠOLSKIH DEJAVNOSTI Center urbane kulture ...

Results for query: "trgovina"
127 results found in 1ms

Freq.	Document	Snippets (limited to 3)
364	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 ...
96	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 ...
92	evem.gov.si.21.html	... eVEM > Področja Trgovina Tu boste našli ... Seznam dejavnosti Druga trgovina na drobno v ... nespecializiranih prodajalnah Druga trgovina na drobno zunaj ...

Results for query: šocial services"
4 results found in 1ms

Freq.	Document	Snippets (limited to 3)
5	e-uprava.gov.si.9.html	... Labour , retirement Social services , health ... relationship etc. ? Social services , health ...
5	e-uprava.gov.si.45.html	... Labour , retirement Social services , health ... relationship etc. ? Social services , health ...
1	podatki.gov.si.340.html	... recreation and spa services ltd . TERME ...

Results for query: "delovno razmerje"
58 results found in 3ms

Freq.	Document	Snippets (limited to 3)
12	evem.gov.si.398.html	... objaviti vsako prosto delovno mesto oziroma vrsto ... , da prosto delovno mesto oziroma vrsto ... zaposlovanje vsako prosto delovno mesto oziroma vrsto ...
8	evem.gov.si.11.html	... razmerja . Delovno razmerje lahko sklenete za ... Najprej objavite prosto delovno mesto , izberete ...
7	evem.gov.si.73.html	... Sloveniji se delovno razmerje načeloma sklepa na ... vstopiti v delovno razmerje . Pogodba o ...

Results for query: "najpogostejši priimki slovenskih prebivalcev"
42 results found in 2ms

Freq.	Document	Snippets (limited to 3)
6	podatki.gov.si.340.html	... SVETA TROJICA V SLOVENSKIH GORICAH Občina Sveti Andraž v Slovenskih goricah OBČINA SVETI ... SVETI JURIJ V SLOVENSKIH GORICAH OBČINA SVETI ... interesno združenje ZDRUŽENJE SLOVENSKIH FESTIVALOV , gospodarsko ...
5	podatki.gov.si.207.html	... , podatki v slovenskih tolarjih in informativni ... , podatki v slovenskih tolarjih in informativni ... , podatki v slovenskih tolarjih in informativni ...

Results for query: "društvo"
30 results found in 1ms

Freq.	Document	Snippets (limited to 3)
164	podatki.gov.si.340.html	... - LEKARNAR BRODARSKO DRUŠTVO STEKLARNA HRASTNIK BROVET ... ŠOLA SLOVENJ GRADEC DRUŠTVO GORSKA REŠEVALNA SLUŽBA KAMNIK DRUŠTVO GORSKE REŠEVALNE SLUŽBE ŠKOFJA LOKA DRUŠTVO KOROŠKI MEDGENERACIJSKI CENTER DRUŠTVO SOŽITJE KAMNIK - DRUŠTVO ZA POMOČ OSEBAM ... DUŠEVNEM RAZVOJU KAMNIK DRUŠTVO SOŽITJE MEDOBČINSKO DRUŠTVO ZA POMOČ OSEBAM ...
9	evem.gov.si.648.html	... oblike podjetij / Društvo Republika Slovenija SPOT ... oblike podjetij > Društvo Društvo Društvo je prostovoljno , ... skupnih interesov . Društvo si samo določi ...