Seminarska naloga 3

Ana Čepuran, Neža Bačar

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

1. Uvod

V tej seminarski nalogi smo implementirali preprost iskalnik po HTML straneh in sicer je bila implementacija sestavljena iz treh glavnih delov:

1. Predprocesiranje in indeksiranje,
2. Poizvedovanje z obrnjenim indeksom,
3. Poizvedovanje brez obrnjenega indeksa.

2. Predprocesiranje in indeksiranje

Predprocesiranje in indeksiranje izvajamo za vsak HTML dokument posebaj. Besedilo dokumentov smo pridobili s pomočjo knjižnice Beautiful Soup, pri čemer smo najprej izločili elemente, kot so: style, script, noscript, head, title, meta in document. Pridobljeno besedilo smo tokenizirali v seznam besed s pomočjo knjižnice NLTK. Vse besede v seznamu smo pretvorili v male tiskane črke in jih primerjali z besedami v seznamu slovenskih stopword-ov, ki smo jih dobili kot gradivo. V primeru, da je beseda na omenjenem seznamu smo jo odstranili iz seznama. Preostale besede smo shranili v seznam. Zgradili smo slovar besed z njihovo frekvenco (število pojavitev besede) in seznamom odmikov (indexi pojavitve). Ta slovar smo nato zapisali v podatkovno bazo SQLite.

3. Poizvedovanje z obrnjenim indeksom (SQLite search)

Na enak način kot smo prej predprocesirali besedilo posameznih HTML datotek, zdaj predprocesiramo željeni poizvedbeni niz – razčlenimo besede, vsako pretvorimo v male črke in preverimo, če se ujemajo s tistimi na seznamu stopwordow. S pridobljenim poizvedbenim nizom nato izvedemo SQL poizvedbo, ki iz tabele Posting vrne vse vrstice, kjer se beseda ujema z eno izmed besed v poizvedbenem nizu. SQL poizvedba nam vrne seznam vrstic, ki ga uporabimo, da zgradimo slovar dokumentov. Za pridobitev besed, ki so okoli iskane besede ponovno odpremo in preberemo HTML dokument, ki vsebuje iskano besedo in s pomočjo seznama odmikov poiščemo tri sosednje besede na vsaki strani. Rezultat poizvedbe izpišemo.

4. Poizvedovanje brez obrnjenega indexa (Basic search)

Poizvedbeni niz se predprocesira na enak način kot v prejšnjih dveh korakih. Potem odpremo vsak dokument v seznamu in ga predprocesiramo na enak način kot smo jih z namen indeksiranja. Rezultat predprocesiranja je slovar besed s številom pojavitev in seznamom odmikov, ki ga sortiramo padajoče glede na število pojavitev besede. Nato na enak način kot pri poizvedovanju z indeksiranjem za generiranje "snippeta" oziroma izpisa, ponovno odpremo HTML datoteko, v kateri se je pojavila beseda in poiščemo sosednje besede ter formatiramo izpis. Pri izpisu smo se odločili ignorirati izpise imen filov, kjer se beseda ne pojavi, oziroma je frekvenca pojavitve enaka 0.

5. Primerjava hitrosti poizvedb

V navodilih je bilo omenjeno, da bo poizvedovanje brez obrnjenega indexa trajalo več časa kot poizvedovanje z obrnjenim indeksom. Pri izvedbi posameznih poizvedovanj smo vključili izpis časa, ki ga metoda porabi za poizvedbo. Ugotovili smo, da poizvedovanje brez obrnjenega indeksa res porabi več časa, kot druga metoda poizvedovanja brez obrnjenega indeksa, seveda pa je to na račun tega, da imamo v prvi metodi že vnaprej shranjene podatke o številu pojavitev in indexih pojavitev. Primerjava porabljenega časa je predstavljena spodaj v tabeli.

METODA POIZVEDBE / POIZVEDBA	Z OBRNJENIM INDEXOM (SQLite Search)	BREZ OBRNJENEGA INDEXA (Basic Search)
predelovanlne dejavnosti	54.23 s	183.31 s
trgovina	15.41 s	128.96 s
social services	2.83 s	126.60 s

Tabela 1: Primerjava trajanja posameznih metod za testne poizvedbe

6. Rezultati

V nadaljevanju so prikazani izbrani smiselni rezultati za željene testne poizvedbe. Vključili smo le prva dva dokumenta, kjer je bila pojavitev besede največja, čeprav sledi veliko število dokumentov, ki vključujejo iskano poizvedbo.

* Izpisi se nadaljujejo, saj je v dokumentu veliko število pojavitev, vendar je izpis predolg, da bi ga v celoti vključili v poročilo.

- **PREDELOVALNE DEJAVNOSTI**

Frequencies Document Snippet

---------- ---------- ----------

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 ... področja C (Predelovalne dejavnosti) predelava ... iskanje ustrezne šifre dejavnosti /storitve in informacij ... pogojih za opravljanje dejavnosti. V iskalnik ... 645 od 645 dejavnosti Izpisanih je od ... Izpisanih je od dejavnosti A KMETIJSTVO IN ... *

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 ... II v zdravstveni dejavnosti Laboratorijski tehnik Ladijski ... Logoped v zdravstveni dejavnosti M Magister farmacije ... funkcije Nosilec obrtne dejavnosti - betoniranje Nosilec ... betoniranje Nosilec obrtne dejavnosti - brušenje in ... varjenjem Nosilec obrtne dejavnosti - električna popravila ... *

- **TRGOVINA**

Frequencies Document Snippet

---------- ---------- ----------

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 ... gl. 46.380 trgovina na drobno s ... Skladiščenje nevarnih kemikalij Trgovina na debelo z ... z nevarnimi kemikalijami Trgovina na drobno z ... gl. 32.500 trgovina na debelo s ... gl. 46.460 trgovina na drobno s ... gl. 38.320 trgovina (odkup in ... gl. 38.220 trgovina na debelo z ... in tehnologijo G TRGOVINA ; VZDRŽEVANJE IN ... MOTORNIH VOZIL 45 Trgovina z motornimi vozili ... *

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 ... z živili Druga trgovina na drobno zunaj ... Nepremičninsko posredovanje Nespecializirana trgovina na debelo Nespecializirana ... na debelo Nespecializirana trgovina na debelo z ...*

- **SOCIAL SERVICES** (izpis je podan v celoti, saj je število pojavitev malo)

Frequencies Document Snippet

---------- ---------- ----------

5 e-uprava.gov.si.9.html ... Labour, retirement Social services, health ... etc. ? Social services, health ... I obtain financial social assistance ? How ..., retirement Social services, health, ... ? Social services, health, ...

5 e-uprava.gov.si.45.html ... Labour, retirement Social services, health ... etc. ? Social services, health ... I obtain financial social assistance ? How ..., retirement Social services, health, ... ? Social services, health, ...

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

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

- **JAVNA USTANOVA**

Frequencies Document Snippet

---------- ---------- ----------

50 e-uprava.gov.si.30.html ... 1. DOL Javna prireditev, organizator ... v daljinskem plavanju Javna prireditev, organizator ... LIVE - ILUZIONISTI Javna prireditev, organizator ... POLETNI ŠPORTNI DNEVI Javna prireditev, organizator ... ZABAVIŠČNI PARK " Javna prireditev, organizator ... TEKME I. LIGE Javna prireditev, organizator ... NKBM v košarki Javna prireditev, organizator ... *

46 podatki.gov.si.340.html ... BRUNO BRESCHI, USTANOVA ZA OHRANJANJE STAREJŠIH ... VAJKARD VALVASOR, USTANOVA ZA PROUČEVANJE IN ... ZA ODVISNIKE - USTANOVA ZA ZDRAVLJENJE ODVISNIKOV ... d.o.o. ITF USTANOVA ZA KREPITEV ČLOVEKOVE ... CENTER PIVKA, USTANOVA ZA POSPEŠEVANJE RAZVOJA ... *

- **MOBILNOST** (izpis je podan v celoti, saj je število pojavitev malo)

Results for a query: "mobilnost"

Frequencies Document Snippet

---------- ---------- ----------

1 e-uprava.gov.si.38.html ... PASS Za popolno mobilnost uporabite prijavo z ...

1 podatki.gov.si.340.html ... REPUBLIKE SLOVENIJE ZA MOBILNOST IN evropske programe ...

- **ELEKTRIČNA PORABA**

Results for a query: "električna poraba"

Frequencies Document Snippet

---------- ---------- ----------

10 podatki.gov.si.9.html ... organi Energetika Končna poraba energije po namenu ... naslovom `` Končna poraba energije ... Nadaljujte ... Državni organi Energetika Poraba energije in goriv ... z naslovom `` Poraba energije in ... ... organi Energetika Energetska poraba goriv, električne ... naslovom `` Energetska poraba ... Nadaljujte z ... Državni organi Energetika Električna energija ( GWh ... z naslovom `` Električna energija ... Nadaljujte ... Državni organi Energetika Električna energija (GWh ... z naslovom `` Električna energija ... Nadaljujte ...

10 podatki.gov.si.496.html ... organi Energetika Končna poraba energije po namenu ... naslovom `` Končna poraba energije ... Nadaljujte ... Državni organi Energetika Poraba energije in goriv ... z naslovom `` Poraba energije in ... ... organi Energetika Energetska poraba goriv, električne ... naslovom `` Energetska poraba ... Nadaljujte z ... Državni organi Energetika Električna energija ( GWh ... z naslovom `` Električna energija ... Nadaljujte ... Državni organi Energetika Električna energija (GWh ... z naslovom `` Električna energija ... Nadaljujte ...