3. seminarska naloga pri predmetu iskanje in ekstrakcija podatkov s spleta

Anže Gregorc, Jaka Kordež

Maj 2019

1 Uvod

Tema tretje seminarske naloge je bila izdelava programa za iskanje po podatkovni zbirki dokumentov. Vsi dokumenti so bili HTML strani iz domen e-prostor.gov.si, e-uprava.gov.si, evem.gov.si in podatki.gov.si.

2 Implementacija

Okolje .NET Core se je že pri prvih dveh nalogah izkazalo za ustrezno, zato sva se zanj odločila tudi tokrat. Uporabila sva knjižnico za dostop do SQLite podatkovne baze in HtmlAgilityPack za obdelavo dokumentov HTML.

3 Indeksiranje

Algoritem za indeksiranje odpre in naloži vsak dokument iz zbirke. Nato iz njega s pomočjo XPath izraza odstrani vse značke tipa *script*, *style* in *noscript*. Nato iz besedila dokumenta z regularnimi izrazi odstrani še številke in večkratne presledke zamenja z enojnimi. Tako prečiščeno besedilo je pripravljeno na pretvorbo v žetone. Tudi ta korak se izvede s pomočjo regularnega izraza. Za vsak žeton, ki ni enak kakšni od stop besed se na koncu vstavi v podatkovno bazo. Indeks žetona v bazi je indeks vključno s stop besedami.

4 Iskanje

Iskalnik najprej poišče vse vrstice v tabeli *Posting*, ki vsebujejo vsaj eno od besed iz iskalnega niza. Te vrstice so nato razvrščene v skupine po dokumentih tako, da dobimo za vsak dokument vsoto pojavitev in skupen seznam indeksov besed iz poizvedbe. Algoritem nato za vsak indeks v vsakem dokumentu doda še predhodnja in sledeča dva, da dobimo zaporedje 7 besed, kjer je iskana na

sredini. Podvojeni indeksi se odstranijo za primer, ko se dve iskani besedi pojavita skupaj. Nato iterira preko vseh žetonov, dobljenih na isti način kot pri indeksiranju in med posamezne odseke z rezultatom vriva tripičja.

Pri iskanju brez uporabe indeksa je postopek povsem enak, le da na začetku ne išče pojavitev v tabeli *Posting*. Namesto tega požene logiko za indeksiranje, ki tokrat išče le besede iz poizvedbe in podatkov ne shranjuje v bazo.

5 Rezultati

Sledijo izmerjeni časi in dobljeni rezultati za nekatere iskalne nize z- in brez uporabe indeksa.

Predelovalne dejavnosti 1902ms 22172ms Trgovina 2030ms 22403ms Social services 1936ms 22053ms Javna uprava 2013ms 21849ms krompir zelje solata korenje koruza 1833ms 21421ms univerza v ljubljani 1911ms 22294ms

Zaradi lepše preglednosti so vsi dobljeni rezultati shranjeni v GitHub repozitoriju v mapi results.