Algoritmi za projekat Search-engine

(OISISI-2019/2020)

Student 1 – Nikola Zejak RA53/2016

**- Ucitavanje direktorijuma i parsiranje html stranica:**

- Kada korisnik odabere opciju unosa direktorijuma iz menija aplikacije, njegov unos se smesta u promenljivu ‘putanja’ koja se proverava da li postoji. Ukoliko ne postoji izbacice poruku. Ako putanja postoji kreirace se tri promenljive: dirpath – putanje do svih direktorijuma na toj putanji, dirnames – njihova imena i filenames – imena fajlova na tim putanjama. Zatim se iterira kroz sve fajlove na datoj putanji i proverava se da li neki od njih ima ektstenziju ‘.html’ ili ‘.htm’. Ukoliko je taj uslov zadovoljen, pokrece se parsiranje html fajlova uz pomoc parser2.py (parser koji je preuzet sa vezbi predmeta). Zatim, kreira se nova promenljiva ‘file’ koja predstavlja spojenu putanju i naziv samog fajla sa ekstenzijom .html. Nakon toga se iterira kroz sve reci(word) koje su isparsirane, a potom se vrsi dodavanje reci I njihovih putanja u stablo.

**- Unos upita za pretragu:**

- Kada korisnik izbere drugu opciju (unos reci za pretragu) prvo se proverava koliko je uneo reci. Prvo se vrsi provera za postojanje logickih operatora and, or i not. Potrebni uslovi za to su da ima tacno tri unute reci od koji je ona na poziciji dva, jedan od operatora. Zatim se vrse provere druge reci, u ovom slucaju koji je logicki operator unet.

1) OR – Na pocetku se proverava da li su prva ili treca rec ‘OR’. Ukoliko jeste, to nije dozvoljeno I izbacie poruku u vidu greske i zamoliti korisnika da ponovo unese reci za pretragu. Ukoliko to nije slucaj, vrsice pretragu datih reci unutar stable I ispisati rezultat.

2) AND – Na pocetku se proverava da li su prva ili treca rec ‘AND. Ukoliko jeste, to nije dozvoljeno I izbacie poruku u vidu greske i zamoliti korisnika da ponovo unese reci za pretragu. Ukoliko to nije slucaj, vrsice pretragu datih reci unutar stable I ispisati rezultat. Napomena: obe reci moraju biti pronadjene. Ako to nije slucaj ispisace koja rec nije pronadjena I zamoliti nov unos reci za pretragu.

3) NOT – Na pocetku se proverava da li su prva ili treca rec ‘NOT’. Ukoliko jeste, to nije dozvoljeno I izbacie poruku u vidu greske i zamoliti korisnika da ponovo unese reci za pretragu. Ukoliko to nije slucaj, vrsice pretragu datih reci unutar stabla i ispisati rezultat. Posto NOT treba da radi pretragu samo prve reci, ukoliko rezultat pretrage prve reci postoji, a rezultat pretrage druge reci je prazan, tek onda ce ispisati glavni rezultat pretrage reci.