Marko Marinković, Tit Arnšek in Damijan Randl

Poročilo

3. del projekta pri predmetu Iskanje in ekstrakcija podatkov s spleta

# Povzetek

V poročilu so predstavljeni rezultati treh postopkov. To so:

* A: indeksiranje podatkov in shranjevanje v bazo
* B: pridobivanje podatkov z inverznim indeksom
* C: pridobivanje podatkov brez inverznega indeksa

# indeksiranje podatkov in shranjevanje v bazo

## Shranjevanje v bazo

Za vsako stran smo najprej z uporabo BeautifulSoupa in WordTokeniza dobili vse besede v besedilu. Vzeli smo samo tekst iz body-ja.

V bazi smo naredili tudi novo tabelo Document s polji:

* documentName
* text
* tokens\_celoten

Iz vseh filtriranih besed(WordTokenize brez StopWords in še nekaterih ostalih, ki smo jih sami dodali) smo določili indekse, na katerih se nahajajo v nefiltriranih tokensih. Tabelo nefiltriranih besed(to so besede v prvotni obliki skupaj z ostalimi ločili in znaki) smo prav tako zapisali v bazo in sicer v tabelo Document, polje tokens\_celoten. To nam je pohitrilo iskanje snippetov v nadaljevanju.

1. pridobivanje podatkov z inverznim indeksom

Tukaj smo za vsako besedo iz query-ja pogledali v katerih dokumentih se nahaja. Za vsak dokument smo tako dobili indekse za vse bedede iz query -ja. Te indekse smo sortirali po velikosti. Če so bili indeksi dovolj skupaj, potem smo več indeksov združil skupaj, da je nastal samo en snippet. Naprimer, če so bili indeksi 5,7,9,11,20, potem smo indekse 5,7,9,11 združili v en snipet.

Za snippet smo uporabil besedo, ki se pojavi na določenem indeksu in še dve besedi levo in desno.

1. pridobivanje podatkov brez inverznega indeksa

Podobno kot smo naredili v točki B, smo storili tudi tukaj le da pri tem nismo uporabljali baze ampak smo pri vsaki poizvedbi pregledali vse dokumente. Pri tem smo uporabili iste metode, kot smo jih uporabili v A in B točki(BeautifulSoup, WordTokenize, Združevanje indeksov… )

# Baza

Baza je sestavljena iz treh tabel:

* IndexWord
* Posting
* Document

Tabeli IndexWord in Posting sta nespremenjeni. Dodali smo tabelo Document zaradi hitrejših poizvedb, saj smo v to tabelo shranjevali tabelo vseh besed skupaj z ločili.

Indeksirali smo 49047 besed. Število bi lahko še zmanjšali, saj sta wordTokenize in BeautifulSoup vrnila besede, ki niso 'smiselne' (stavbi.za, 2017poročilo, 'izvoz, …)

Beseda, ki se največkrat pojavi v enem dokumentu je proizvodnja: v dokumentu evem.gov.si.371.html se pojavi 2266-krat.

Vsaj enkrat v dokumentu se pojavi največkrat beseda 'pogoji'

Slika, ki vsebuje besede besedilo, posnetek zaslona, številka, pisava

Opis je samodejno ustvarjen

Beseda, ki se največkrat pojavi v vseh dokumentih skupaj:

Slika, ki vsebuje besede besedilo, posnetek zaslona, številka, pisava

Opis je samodejno ustvarjen

Dokument z največjo vsoto frekvenc:

Slika, ki vsebuje besede besedilo, posnetek zaslona, pisava, številka

Opis je samodejno ustvarjen

## Outputs

### Overstock.com

### Komentar:

Uporaben del tega izhoda je označen z rdečo barvo. Ta del predstavlja posamezne izdelke na spletni strani, kjer se lepo opazi, da se spreminjo ključni podatki.

# Viri

* Parsel [Online]. Dosegljivo: <https://parsel.readthedocs.io/en/latest/usage.html> (zadnji dostop 5.5.2023)
* Beautiful Soup Documentation [Online]. Dosegljivo: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/> (zadnji dostop 5.5.2023)
* Python RegEx, W3Schools [Online]. Dosegljivo: <https://www.w3schools.com/python/python_regex.asp> (zadnji dostop 5.5.2023)
* V. Crescenzi, G. Mecca, P. Meraildo. ROADRUNNER: Towards Automatic Data Extraction from Large Web Sites [Online]. Dosegljivo: <https://vldb.org/conf/2001/P109.pdf> (zadnji dostop 5.5.2023)
* Feature Film, Rating Count at least 25,000, Action (Sorted by IMDb Raiting Descending) [Online]. Dosegljivo:
* <https://www.imdb.com/search/title/?genres=action&sort=user_rating,desc&title_type=feature&num_votes=25000,&pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=94365f40-17a1-4450-9ea8-01159990ef7f&pf_rd_r=VTPTT29EC5M36Q86ZTPK&pf_rd_s=right-6&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_gnr_1> (zadnji dostop 5.5.2023)
* Feature Film, Rating Count at least 25,000, Action (Sorted by IMDb Raiting Descending) [Online]. Dosegljivo:
* <https://www.imdb.com/search/title/?genres=sport&sort=user_rating,desc&title_type=feature&num_votes=25000,&pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=94365f40-17a1-4450-9ea8-01159990ef7f&pf_rd_r=VTPTT29EC5M36Q86ZTPK&pf_rd_s=right-6&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_gnr_18> (zadnji dostop 5.5.2023)