Marko Marinković, Damijan Randl in Tit Arnšek

Programming assignment 1

Seminarska naloga pri predmetu Iskanje in ekstrakcija podatkov s spleta

Uvod

Za prvi projekt pri predmetu smo morali implementirati spletnega pajka, ki iz podanih semenskih strani najde in shrani podatke o vseh slikah, dokumentih in linkih. Obišče tudi vse linke, ki jih najde, pri tem pa se omejimo zgolj na strani iz domen z obliko *gov.si.

Implementacija

Naš pajek podpira delovanje z več niti.
Ko pajek dobi naslov iz frontierja, najprej preveri, ali že obstaja domena(to samo pri semenskih straneh, saj je v nadaljevanju ta pogoj že zagotovljen, zato preverjanje domene preskoči). Če se je domena pojavila prvič, potem domeno doda v bazo skupaj s podatki o robots.txt in sitemap. V naslednjem koraku pajek preveri ali že lahko dostopa do strani. Ta pogoj preveri tako, da v tabeli site, shranjuje zadnje dostope do domene. Tako je vedel, ali je že preteklo dovolj časa od zadnjega dostopa do strani s to domeno.
Nadaljuje z obdelavo strani:

- HTTP status koda
- Pridobivanje vsebine iz strani(html content)
- Pridobivanje linkov
- Pridobivanje slik
- Pridobivanje datotek

Zgornje podatke smo dobivali s pomočjo knjižnic Requests in Selenium.
Preden začne pridobivati linke, slike in datoteke, preveri ali je podana stran duplikat. To preveri tako, da pogleda, ali že obstaja kakšna stran v bazi z enako vsebino. Če ne gre za duplikat nadaljuje z obravnavo.
Pridobljene linke doda v bazo(kot FRONTIER), hkrati pa za vsak link še preveri njegovo domeno in v primeru, da je še ni v bazi, jo doda. V bazo je dodal tudi linke, ki niso iz domene *gov.si, vendar je le te označil kot ZUNANJI in ne kot FRONTIER. Zunanjih linkov v nadaljevanju ni obdeloval. Linke, ki je dodal v bazo kot FRONTIER, doda tudi v vrsto frontier, iz katere je pajek pridobival naslednje strani.
Za slike, ki jih je pridobil, je najprej preveril, ali so ustreznega formata. Te je nato dodal v bazo, in sicer v tabelo page kot BINARY, prav tako pa v tabelo image.
Podobno je storil tudi za datoteke, shrani jih v tabelo page kot BINARY in v tabelo page_data.
Vse zgoraj omenjene linke, strani in datoteke hkrati shranjuje tudi v tabelo link.

- Funkcionalnost, povezana s pridobivanjem podatkov iz strani, je implementirana v razredu Page in Vmesnik.
- Funkcionalnost, povezana s pridobivanjem in vstavljanjem podatkov v bazo, je implementirana v baza.py.
- Funkcionalnost, povezana z domenami, je implementirana v robots.py.

Problemi

Pri obdelavi strani, smo sproti iskali in popravljali napake, zato so v bazi nekonsistentni podatki.

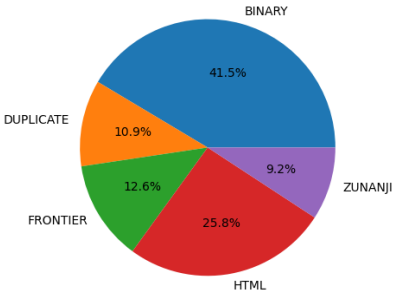
- Da pajki niso hkrati dostopali do iste domene, smo v tabelo site, dodali tudi polje zadnji_dostop. Tako smo lahko zagotovili, da nikoli nismo dostopali do iste domene prepogosto.

- Določenih linkov na začetku nismo uspeli ujeti, saj smo šele kasneje ugotovili, da nismo upoštevali vse možnosti, kjer bi se linki lahko skrivali.
- Podoben problem je bil tudi pri slikah in datotekah, zato obstaja možnost, da nam v bazi manjka nekaj datotek in slik.
- Nekaj težav je bilo tudi z implementacijo večnitnega pajka, saj smo na začetku delali zgolj z enim.
- Pri macOS je Selenium driver prenašal datoteke.
- Večnitnega pajka smo poskušali poganjati z različnimi števili niti. Pri povečanju števila se hitrost pridobivanja strani ni bistveno povečala, saj so bili linki na nekem odseku iz frontierja pogosto iz iste domene, kar je onemogočalo, da bi več pajkov hkrati obdelovalo isto stran.

Statistika

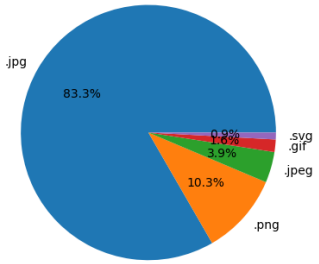
Do sedaj je pajek v približno 65 urah delovanja v bazo shranil 76248 strani.
V tabeli je prikazano število posameznih strani glede na type code.

PAGE TYPE	ŠTEVILO
BINARY	31615
HTML	19691
FRONTIER	9633
DUPLICATE	8305
ZUNANJI	7004



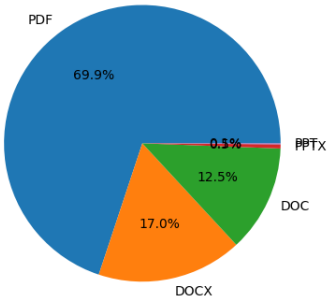
Slike deljene glede na format:

CONTENT TYPE	ŠTEVILO
JPG	13141
PNG	1627
JPEG	616
GIF	249
SVG	137



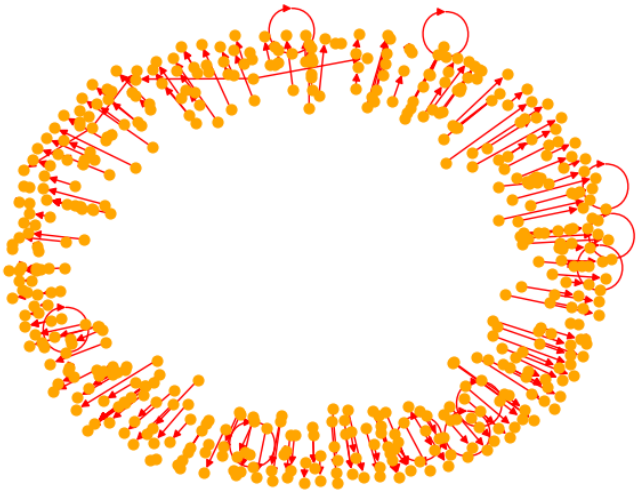
Datoteke deljene glede na format:

DATA TYPE	ŠTEVILO
PDF	11116
DOCX	2703
DOC	1989
PPTX	77
PPT	18

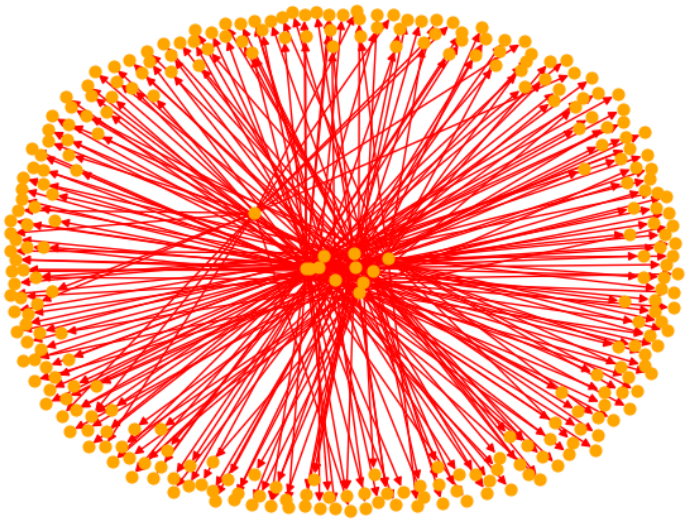


Domene deljene glede na semenske strani:

DOMENA	ŠTEVILO
gov.si	30497
e-uprava.gov.si	1912
e-prostor.gov.si	180
evem.gov.si	147
OSTALO	43512



Slika 1: Prikaz naključno izbranih 300 povezav



Slika 2: prikaz strani z največ outlinki