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 večimi 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 dodamo v bazo skupaj s podatki o robots.txt in sitemap. V naslednjem koraku je pajek preveril ali že lahko dostopa do strani. Ta pogoj smo preverili tako, da smo v tabeli site, hranili zadnji dostop do domene. Tako je pajek vedel, ali je že preteklo dovolj časa od zadnjega dostopa strani z to domeno.

Nadaljujemo 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 začnemo pridobivati linke, slike in datoteke, preverimo ali je podana stran duplikat. To preverimo tako, da pogledamo, ali že obstaja kakšna stran v bazi z enako vsebino. Če ne gre za duplikat nadaljujemo z obravnavo.

Pridobljene linke dodamo v bazo(kot FRONTIER), hkrati pa za vsak link že preverimo njegovo domeno in v primeru, da je še nimamo v bazi, jo dodamo. V bazo smo dodali tudi linke, ki niso iz domene \*gov.si, vendar smo le te označili kot ZUNANJI in ne kot FRONTIER. Zunanjih linkov v nadaljevanju nismo obdelovali. Linke, ki smo dodali v bazo kot FRONTIER, dodamo tudi v vrsto Frontier, iz katere so pajki pridobivali naslednje strani.

Za slike, ki smo jih pridobili, smo najprej preverili, ali so ustreznega formata. Te smo nato dodali v bazo, in sicer v tabelo page kot BINARY, prav tako pa v tabelo image.

Podobno smo storili tudi za datoteke, shranili smo jih v tabelo page kot BINARY in v tabelo page\_data.

Vse zgoraj omenjene linke, strani in datoteke smo hkrati shranjevali tudi v tabelo link.

Funkcionalnost, povezana s pridobivanjem podatkov iz strani, je implementirana v razredu Page.

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

**Slika, ki vsebuje besede diagram, tortni grafikon

Opis je samodejno ustvarjen**Do sedaj je pajek v bazo shranil NEKAJ strani. V tabeli je prikazano število posameznih strani glede na type code.

|  |  |
| --- | --- |
| PAGE TYPE | ŠTEVILO |
| BINARY |  |
| HTML |  |
| FRONTIER |  |
| ZUNANJI |  |
| DUPLICATE |  |

Slike deljene glede na format:

|  |  |
| --- | --- |
| CONTENT TYPE | ŠTEVILO |
| JPG |  |
| PNG |  |
| JPEG |  |
| GIF |  |
| SVG |  |

Datoteke deljene glede na format:

|  |  |
| --- | --- |
| DATA TYPE | ŠTEVILO |
| PDF |  |
| DOCX |  |
| DOC |  |
| PPTX |  |
| PPT |  |

Domene deljene glede na semenske strani:

|  |  |
| --- | --- |
| DOMENA | ŠTEVILO |
| gov.si |  |
| e-uprava.gov.si |  |
| e-prostor.gov.si |  |
| evem.gov.si |  |
| OSTALO |  |

Slika, ki vsebuje besede krog

Opis je samodejno ustvarjen

Slika : Prikaz naključno izbranih 300 povezav

**Slika, ki vsebuje besede nevretenčar

Opis je samodejno ustvarjen**

Slika : prikaz strani z največ outlinki