# Poročilo projekta Webcrawley (spletnega pajka)

Aljaž Rupar, Gregor Ažbe, Nal Lukšič

## 1 Uvod

V tem poročilu bomo predstavili končni produkt razvoja spletnega pajka, ki brska po straneh slovenske vlade. To so spletne strani, ki spadajo pod domeno gov.si. Pajek upošteva osnovna etična načela brskanja po straneh in učinkovito vrača podatke iz spletišč. Pridobljeni podatki so nato shranjeni v lokalno vzpostavljeni podatkovni bazi. Avtorji projekta smo Nal Lukšič, Aljaž Rupar in Gregor Ažbe.

## 2 Specifikacije pajka

Za implementacijo spletnega pajka smo uporabili programski jezik Python. Za zajem spletne strani smo uporabili knjižnico Urllib, za razčlenjevanje pa BeautifulSoap.

*Frontier* smo implementirali kot vrsto. Za to smo uporabili Pythonov vgrajen podatkovni tip *Queue.*

Sočasnost smo implementirali s pomočjo niti. Izbrali smo med uporabo niti in procesov. Uporaba več niti v implementaciji programskega jezika Python sicer ne omogoča sočasnega izvajanja niti. Kljub temu pa smo uporabili večnitnost, saj je za spletne pajke specifično, da imajo veliko čakanja (npr. pridobivanje spletnih strani, pisanje v bazo, ipd.). Poleg tega pa imajo niti skupen naslovni prostor kar je olajšalo uporabo globalnih spremenljivk. Prav zaradi tega pa je bilo potrebno poskrbeti še za sinhronizacijo med nitmi pri uporabi skupnih virov. To smo dosegli s ključavnicami, ki so prav tako del programskega jezika Python.

Naslednji problem, ki se je pojavil pri uporabi niti, pa je bil zagotavljanje časovnega razmika med posameznimi zahtevami na isti IP naslov. To smo dosegli tem, da smo za posamezen IP naslov hranili čas zadnjega dostopa. Če od zadnjega dostopa še ni minilo zahtevano število sekund, nit počaka pet sekund spet preveri, če že lahko pošlje zahtevo za določeno stran. Če pa je od zadnjega dostopa že minilo zahtevano število sekund, se kot zadnji dostop zabeleži trenutni čas in prične z obdelavo naslova.

### 2.1 Parametri in posebnosti

### 2.2 Ovire pri razvoju

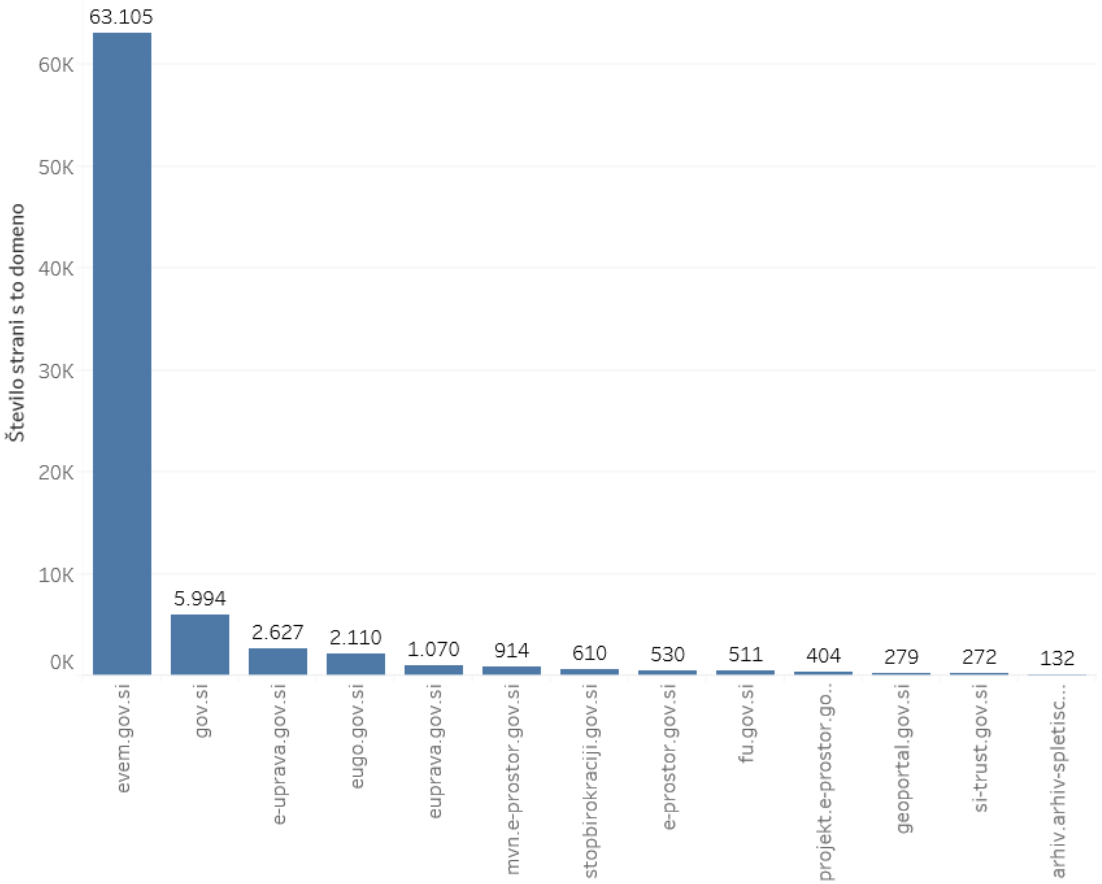
Eden od problemov na katerega smo naleteli, je bilo ustavljanje niti zaradi napak. Če je pri obdelovanju spletne strani prišlo do napake je je niz ustavila. Tako je bilo delovanje vedno

počasnejše, dokler se ni popolnoma ustavilo. To smo rešili tako da smo dodali lovljenje vseh napak pri obdelovanju strani. Tako se je napaka le izpisala, nit pa je nadaljevala z naslednjo stranjo.

Eden od problemov je bil tudi v tem, da se je z vsako ustavitvijo izbrisala vrsta in množica obiskanih strani. To smo rešili tako, da smo poleg shranjevanja v pomnilniku ti podatkovni strukturi shranili tudi v podatkovni bazi.

# Statistike

## 2.3 Vizualna podoba spletišča




## 2.4 Statistika

| Crawldb | | Seed urls | |
|---|---|---|---|
| Number of sites | 176 | Number of sites | 4 |
| Number of web pages | 79890 | Number of web pages | 71850 |
| Number of duplicates | 1162 | Number of duplicates | 810 |
| Number of binary documents - all | 3275 | Number of binary documents - all | |
| Number of binary documents - PDF | 2277 | Number of binary documents - PDF | 1626 |
| Number of binary documents - JPEG | 469 | Number of binary documents - JPEG | 310 |
| Number of binary documents - MSWORD | 317 | Number of binary documents - MSWORD | 269 |

| | | | |
|---|---|---|---|
| Number of binary documents - PNG | 85 | Number of binary documents - PNG | 19 |
| Number of binary documents - TIFF | 25 | Number of binary documents - TIFF | 23 |
| Number of binary documents - TEXT PLAIN | 20 | Number of binary documents - TEXT PLAIN | 19 |
| Number of binary documents - TEXT/XML | 10 | Number of binary documents - TEXT/XML | 9 |
| Number of binary documents - GIF | 5 | Number of binary documents - GIF | 4 |
| Number of binary documents - APP/XML | 4 | Number of binary documents - APP/XML | 0 |
| Number of binary documents - VIDEO | 4 | Number of binary documents - VIDEO | 0 |
| Number of binary documents - RTF | 1 | Number of binary documents - RTF | 1 |
| Number of images | 449935 | Number of images | 409569 |
| Average number of images per web page | 7 | Average number of images per web page | 7 |