Implementacija spletnega pajka

Skupina **DOMACI-NJOKI**

***Abstract*—V seminarju bomo opisali našo implementacijo spletnega pajka.**

I. Uvod

Pajek je razdeljen na več medsebojno povezanih komponent.

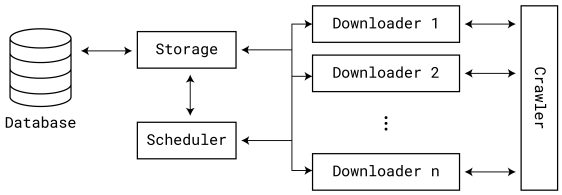

Fig. 1. Arhitektura implementiranega spletnega pajka

A. Razvrščevalnik

Razvrščevalnik (angl. scheduler) URL naslovov je implementiran kot podatkovna vrsta, do katere lahko obenem dostopa več različnih niti.

Ob dodajanju novega URL naslova v vrsto, razvrščevalnik preveri, ali je bil ta URL naslov že prenesen. Prav tako zna iz URL naslova prebrati domeno, pridobiti IP naslov strežnika s to domeno in prenesti datoteko `robots.txt`.

B. Pajek

Ob zagonu, pajek zažene več niti, vsaka od njih pa ima nalogo, da iz razvrščevalnika URL naslovov pridobi naslov, prenese stran in shrani podatke v podatkovno bazo. Pridobivanje vsebine strani izvajamo v dveh korakih:

- Najprej na strežnik izvedemo `HTTP HEAD` metodo.
- S pomočjo Selenium brskalnika obiščemo stran.

C. Shramba

II. Zaključek