Järeldused ja tulemused

Herman Karu

Spideri loomine oli minu jaoks esmakordne, abiks pidin võtma internetis oleva info, millest lähtuvalt sai loodud antud spider. Spideri pidin mitu korda ümber kirjutama, sest ei saanud sobivat tulemust, seega kulus selle jaoks kahe päeva jooksul 4-5 tundi. Lõpptulemus ei ole täpselt see, mida ma tahtsin, sest *output.txt* failis esineb url duplikaate erinevate HTTP päistega ning failis ei ole erinevate päiste hulk täpselt sama, mis tuli Scrapy logifailiga.

Spider on loodud Pythonis kasutades Scrapy raamistikku. Scrapy saab installida Powershelli kaudu kasutades *pip install scrapy* käsklust. Koodi jooksutamiseks tuleb minna Powershelli, otsida üles kaust, milles on Spider\_uus.py fail ning kasutada käsklust *scrapy runspider .\Spider\_uus.py*

Kodus jooksutatud koodi puhul panen lähtekoodiga kaasa *output.txt* faili koos *scrapy\_log.txt* failiga, milles on välja toodud Scrapy statistika. Sellest statistikast lähtudes sai spider enim „GET“ päiseid (3781), teised päised (CONNECT, OPTIONS, PATCH, POST ja TRACE) esinesid 947 korral, kuid HEAD päis on statistikas 950 korral. GET päist võib esineda enim lähtudes Scrapy eripärast. Kokku tegi spider 9466 requesti, millest 6610 said HTTP 200 vastuse ehk loa. Palju esines ka 405 vastuseid ehk server ei lubanud spideril teha konkreetseid muudatusi. Samuti esines HTTP 400 errorit üle tuhande korra. Sellest lähtuvalt saan aru, et rik.ee domeenis on andmekogujatel lubatud üsnagi palju teha, kuid teatud osad domeenist on sellistele kogujatele keelatud.

Koodi väljundiks olevas *output.txt* failis on väikesed erinevused võrreldes Scrapy poolt antud logifailiga, sest väljundfaili on lubatud panna ainult domeeniga seotud toimingud (Scrapi logib ka katsed minna alamdomeeni või konkreetsest domeenist välja). Avastasin, et [www.rik.ee](http://www.rik.ee) kasutab OPTIONS, PATCH, POST, HEAD päiseid, kuid mitte GET ja näiteks CONNECT päiseid. Samas kasutavad paljud teised lingid domeenis OPTIONS, POST, HEAD, GET, PATCH päised. Spiderit tehes avastasin, et kui urlile anda esmalt GET meetod, siis spider leiab üles tavalingid (uudiseid, teadaanded jne), kuid kui alustada HEAD meetodiga, siis spider liigub kiirelt teisejärgulistele lehtedele (printlehed, .pdf faile sisaldavad urlid jne).