Podstawy Teleinformatyki

Web scraper

Daniel Pawlikowski
Szymon Królikowski
Anna Zdrojewska

Prowadzący:
mgr inż. Przemysław Walkowiak

Dlaczego akurat ten projekt ?

Web scraping jest to technika wydobywania oraz gromadzenia interesujących nas informacji dostępnych za pośrednictwem stron internetowych. Sposobem pozyskiwania tego typu danych jest parsowanie kodu HTML wybranych stron za pośrednictwem wcześniej stworzonych wyrażeń regularnych.
Nasz projekt opierać się będzie na pozyskiwaniu licznych informacji zawartych na publicznych profilach w serwisie Instagram tj. zdjęcia, tagi itd. oraz wykorzystaniu ich do celów badawczych. Web scraping posłuży nam również jako sposób obejścia narzuconych przez instagram ograniczeń, które spotkalibyśmy w przypadku próby wykorzystania API instagramowego.
Projekt ten zainteresował nas głównie, ponieważ w dzisiejszych czasach, kiedy coraz więcej informacji dostępnych jest w internecie, web scraping zdobywa na popularności oraz ze względu na fakt iż jest to ważna technika przy zadaniach typu big data i warto byłoby się zaznajomić z tematem.

Przewidywany stack technologiczny:

- język programowania Java
- biblioteka jSoup do parsowania stron
- Selenium do “poruszania” się po stronie
- Git
- Maven/Gradle
- inne technologie, które okażą się potrzebne na drodze rozwoju projektu

Wstępny podział ról w zespole:

- Daniel Pawlikowski, Szymon Królikowski - prace badawcze związane z mechaniką działania serwisu, manipulacją linkami. Opracowanie odpowiednich regexów, implementacja funkcjonalności
- Anna Zdrojewska - wykorzystanie frameworku Selenium w kodzie aplikacji do celów automatyzacji ruchów w przeglądarce, implementacja funkcjonalności