INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER (UOW)

IceScraper – Optimized Data sets provider

A dissertation By

Dulika Lavanya Ranasinghe

Supervised By

Mr. Pumudu Fernando

Submitted in partial fulfillment of the requirements for the BSc (Hons) Software Engineering degree

Department of Computing

May 2018

Abstract

Data is said to be the new oil for the coming decade. Every organization is after the data these days spending millions of dollars to be ahead in the game. Web scraping and bots are among one of the novel tools that businesses are focusing on. Commercial web scraping tools are every where. But there are lot of limitations while it can be used for a given scenario.

Scraping from a single page & creating a web scraping agent are completely two different tasks. Scraping a single page is just an attribute of web scraping agent. Web scraping agent consist of techniques such as by pass IP blocks, deal with bandwidth & Ram , css changes etc. Web scraping agent should be able to scrape multiple pages by handling java script without failures , do that job in speed of single page scrapers' speed, keep the efficiency & consistency. The proposed solution is a framework to scrape multiple web pages with an increased accuracy while maintaing the web elements like Javascript and CSS. The prototype presented would be gathering data sets for a given set of URLs by an user.