Ryan Gilmore, Aidan Sullivan, Lo-Badal Burch

COMP 2100-13
Group 15
Final Project Proposal
Web Scraping Program

Project Overview

We are going to be implementing a web scraping application using the beautiful soup python library. The goal of the program is to collect and process data from websites without the use of csv files or an application programming interface. The program will request web servers using a provided URL. The servers will respond with html files representing the user side of the website using hypertext transfer protocol. The program will then parse the html files, select relevant pieces of data, and write the data to a text file. The web scraping process will work with static web pages but ideally our program will be able to handle websites that send javascript files tailored to users.

Importance/Interest in Web Scraping

The web is an infinite source of information. Data in finance, healthcare and economics is publicly available and easily accessible thanks to websites. While the data is there it is time consuming to surf through dozens of web pages and read through entire spreadsheets. As programmers it is our job to automate processes that can be automated. Databases stored in csv format are easy to retrieve and manipulate through an application programming interface, but what about websites made up of walls of text? This is the problem we hope to solve with our program. With web scraping, we could expedite our search for CO-OPs. We could parse web pages containing job posting to help narrow the positions applicable to us. In addition, the

members of our group are interested in convolutional neural networks. Web scraping allows us to build large data sets to serve as inputs for machine learning projects.

Staged Plan for Implementation

Our project goals will be broken down into a week by week basis, meeting on Wednesday each week to discuss our progress. Additionally we will be using a github repository as source control between group members. We have roughly four and a half weeks to implement our completed project and our presentation. The first week will be spent reading through the beautiful soup library and understanding our data source.We will need to understand which functions we should be implementing in our program and how the html data we will be receiving is formatted. To start we will be using our program to scrape Glassdoor, so understanding the Glassdoor webpage source will be important. The following week will be focused on scraping through the html data. Functions for parsing the data are provided by the beautiful soup library. We will have to supply our program with search fields and determine which pieces of data gathered from the parse are worth storing. Within two weeks we will have a “blueprint” of a web scraping program. At that point our program will be adapted to scrape data from a single provided url. In the following week we will focus on adapting our program to make multiple HTTP requests. This will require some slight modifications to our python script and some drastic changes to the way we search for data on an html file. In the final week we will be working on the presentation and testing our program.