Joanna Wu

January 2020

Advisor: James Glenn

CPSC 490 : Senior Project Thesis Proposal

Python Scripting Applications at Yale: Web Scraping

A primary personal and academic goal for myself with my senior project is to learn a new skill. I decided I wanted to do a deep dive and investigation into the power and ability of Python scripting, specifically in its web scraping applications.

My overall plan for the project is to begin by building a web scraping program that will serve the needs of my a cappella group. Currently, a task that requires a lot of labor and time from the group is tour managing, or contact potential performance hosts in a given city or area. Often times those hosts are schools, and the contacts we are looking for are principals, or music teachers. The fastest way to get this information right now is to google schools in the given area, go to each website individually, search for a principal's email address or a faculty directory, and copy and paste the email information into an excel spreadsheet.

This first part of the project will serve as a proof of concept, to be able to be expanded to further applications. I plan on reaching out to Yale libraries, administrative

offices, department offices, and other organizations to see if they might have similar menial or repetitive tasks which could be accomplished with web scraping. I hope that this will also allow me to learn more about the repeatability of these projects, and how domain differences change the needs of the scraping service.

I will be using online resources like KhanAcademy, Youtube Tutorials, and other coding bootcamps to learn more about Python scripting and web scraping. I'll be using the Python library BeautifulSoup, text editor Sublime, and will host any websites necessary using Heroku.

One potential problem I anticipate with this project is the ability to find a website that is suitable for scraping, i.e. one in which most data is formatted similarly and there are few edge cases. However, after researching I discovered that many public schools in the Washington D.C. area are compiled on the District of Columbia Public Schools page for schools profiles: http://profiles.dcps.dc.gov/ My original proof of concept script will aim to scrape the school names, principal names and principal's email addresses from this database, and export it to a csv file.

Timeline, Deliverables and Deadlines:

January 30	Edit proposal and begin research
Feb 6	Submit proposal to DUS

Feb 20	Research write up on web scraping and python scripting
March 5	Write proof of concept script for Washington DC schools
March 19	Work on edge cases and troubleshooting, reach out to Yale administration
April 2	Determine 1-2 Yale affiliated projects to work on, meet with representatives to determine their needs
April 16	Deliver to administration scripts, work on final project write up
April 30	Senior project due

Final list of deliverables:

- Research write up on Python Beautiful Soup library and web scraping techniques
- Initial proof of concept script for Washington DC public schools
- Report and recommendations for Yale affiliated offices requests for scripts, their needs
- Script written for a Yale affiliated office or administrative service
- Final project write up