CS 5010 Homework 3: Python and Web Scraper

# **Headline Scraping from Major News Sites and Word Frequency Analysis**

## H. Diana McSpadden

## hdm5s

September 21, 2020

Overview

I performed web scraping from six news sites for news headlines over four days. I collected headlines, URLs, datetime of the scraping, and website. Using AllSides Media Bias Chart (AllSides, n.d.) I categorized the headlines as center, lightblue, or lightred. After collecting 1,616 headlines, in a Jupyter Notebook I retrieved, cleaned, and performed word frequency analysis on the data.

Approach

I was interested in scraping multiple sites because I wanted to experience the challenges of python web scraping as the html structure of a scraped website changes. I was also interested in scraping words, not numbers because I wanted to work with some natural language processing in python. I decided to scrape headlines from major U.S. news sources because generally headline text is simple text using few “filler” or “stop” words, and theoretically, there should be alignment between the news topics between major news sites.

In order to add filtering functionality to the assignment I decided to categorize the news sites as center, left-leaning, or right-leaning. I used AllSides.com’s Media Bias Chart (AllSides, n.d.). The news sites I selected are outlined in red in Figure 1.

A screenshot of a cell phone screen with text

Description automatically generated

**Figure 1: AllSides Media Bias Chart with selected news sites circled in red.**

BeautifulSoup 4 python library to parse the structured html content (Real Python, 2020).

Utility

* reasons why someone might want to use your program (utility of the data you scraped - this can be simple/straightforward)

Wish List

* (optional) ways your program could be extended to include additional features

Extra Credit

* (optional) any extra functionality you added to the program beyond the given requirements, as well as any additional results you obtained as a result of it (if appropriate). Call this section “Extra Credit.”

References

<https://pythonprogramming.net/tokenizing-words-sentences-nltk-tutorial/>

AllSides. (n.d.). *AllSides Media Bias Chart* [Illustration]. AllSides. <https://www.allsides.com/sites/default/files/AllSidesMediaBiasChart-Version2.jpg>

*Beautiful Soup - Navigating by Tags - Tutorialspoint*. (n.d.). TutorialsPoint. Retrieved September 14, 2020, from <https://www.tutorialspoint.com/beautiful_soup/beautiful_soup_navigating_by_tags.htm>

Real Python. (2020, August 21). *Beautiful Soup: Build a Web Scraper With Python*. https://realpython.com/beautiful-soup-web-scraper-python/#part-3-parse-html-code-with-beautiful-soup