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 have interest in scraping multiple sites because I wanted to experience the challenge of python web scraping varied html structures. I focused on scraping words, not numbers, because I want to work with some natural language processing in python. For the assignment I scraped headlines from major U.S. news sources. I used headlines because 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 because news is topical. I used the BeautifulSoup 4 python library (Real Python, 2020). to parse the structured html content of the homepages for the following news sites in Table 1.

I also have interest in filtering the resulting data set. For the assignment I included filtering functionality by categorizing the news sites as center, left-leaning, or right-leaning. I used AllSides.com’s Media Bias Chart (AllSides, n.d.), Figure 1, to determine the category for each of the six news sites. The news sites and their categories are listed in Table 1.

|  |  |  |
| --- | --- | --- |
| News Site | URL Used For Scraping | AllSides.com Category |
| ABC News | <https://abcnews.go.com/> | lightblue |
| AP News | <https://apnews.com/> | center |
| CBS News | <https://www.cbsnews.com/> | lightblue |
| Fox News | <https://www.foxnews.com/> | lightred |
| NPR News | <https://www.npr.org/sections/news/> | center |
| Reason | <https://reason.com/> | lightred |

Table 1: Scraped News Sites with AllSides.com Category

I also have interest in language processing and added this dimension to the assignment by utilizing the Natural Language Toolkit library, NLTK (*Tokenizing Words and Sentences with NLTK*, n.d.) to remove stop words, and analyze word frequency in the complete data set, and in the data set filtered by center, left-leaning, or right-leaning categories of news.

A screenshot of a cell phone screen with text

Description automatically generated

Figure : AllSides Media Bias Chart with Selected Sites Circled

Utility

With additional data, over longer time periods, and with a consistent number of records for each of the categories, I believe my web scraping and headline word analysis can identify where left-leaning, center, and right-leaning news coverage diverges, or converges. Identifying words frequently used by a category of news sites, and not by other news site categories can identify blind-spots in coverage, or topics that are over-covered.

Wish List

With more time I would have liked to have added several more news sites to my data set. Each news site has an entirely different html structure, and I found it difficult to reuse my code between websites. Because I wasn’t able to add more websites to my data set, the data set contained many more lightblue records than the other categories. Table 2 shows the number of headlines by category.

|  |  |
| --- | --- |
| Category | Number of Headlines |
| center | 134 |
| lightred | 195 |
| lightblue | 339 |

Table 2: Number of headlines by Category

With more time I also would have liked to investigate the type of words that were in a similar ratio of headlines for each of the categories, and words that were over or under represented in each of the categories.

Extra Credit

My use of the NLTK library to create word frequency plots added an extra dimension to this project. I enjoyed trying to determine how to massage my data frame of headlines to remove stop words, and create valid data for the NLTK to process. I needed to repeated use map(), split(), and join() to format my data, and I unexpectedly found a use for reduce(), which I used to turn a series into a single string.

References

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>

*Tokenizing Words and Sentences with NLTK*. (n.d.). PythonProgramming. Retrieved September 15, 2020, from https://pythonprogramming.net/tokenizing-words-sentences-nltk-tutorial/