Shariq M Jamil

Graduate Project

For this project, I had planned on scraping posts on bogleheads.org for Amazon links. I wrote a scraping engine (using Scrapy) that iterates through each page of the ~250 threads on the Bogleheads front page, storing any external links and information about the containing thread in a local PostgreSQL database. I then wrote a bash script so that the python program could be run using the Linux crontab functionality, which allowed the program to run every five minutes while my laptop was running. As of now, I have 24,437 links stored in the database. Based on initial goals, the data collection phase of this project has gone smoothly and seems to be robust as long as the website's structure remains the same.

For presenting insights gained from collecting these links, I compiled a Jupyter Notebook. I preferred this method because the notebook can run SQL queries and redraw graphs based on new data. I was surprised by how varied and well distributed the data was. There were some interesting insights, as shown in Figure 1 below, Youtube appears as 'www.youtube.com' and 'youtu.be', which implies that it dominates as an information-sharing platform.

I am looking forward to continuing to build on this project by adding new functionalities. I would like to connect to Amazon's API so I can import product data. I would also like to collect the timestamp of when links were posted to be able to plot links against time, and the username of the poster to see if certain users are pushing specific products on the Bogleheads platform.

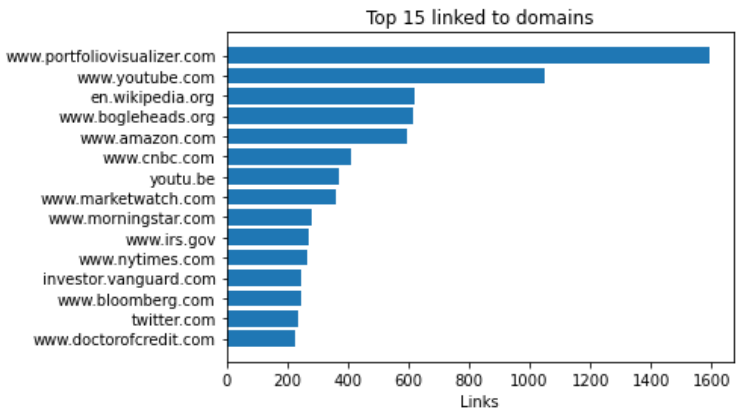

Figure 1: Most linked-to domains