Shariq M Jamil

Graduate Project Checkpoint

For the graduate project, I had planned on scraping posts on bogleheads.org for Amazon links. I have completed writing the scraping engine that iterates through each page of the 250 threads on the Bogleheads front page, storing any external links in a CSV file. Based on over nine thousand links collected, the top five linked domains are as follows:

[('www.portfoliovisualizer.com', 1068), ('www.youtube.com', 697), ('en.wikipedia.org', 280), ('youtu.be', 221), ('www.amazon.com', 219),…]

There are plenty of Amazon links, so I am moving forward with the proposal but, I may analyze other domain links as well (most linked Youtube videos and Wikipedia articles). This data may lead to a crafted index of highly recommended reading or watching material from the forum. To view and manipulate these links, I have set up a PostgreSQL database (Figure 2, 3) and am planning on redirecting data there instead of the CSV file (Figure 1).

After populating the database, I will start working on a Jupyter Notebook to present the data. I will use Pandas to manipulate and visualize link data. As time and data permits, I plan on presenting the following:

- Top 10 Wikipedia and Youtube links.
- Top 10 linked Amazon products
- Top 10 domains
- Integrating with the Amazon API and filtering links by product category.
- Analyzing the data for other findings

```
https://www.amazon.com/Balanced-Asset-Allocation-Economic-Climate/dp/1118711947,328869,https://www.bogleheads.org/forum/viewtopic.php?t=328869,David F. Swensen,www.amazon.com,
```

Figure 1: Sample CSV file entry

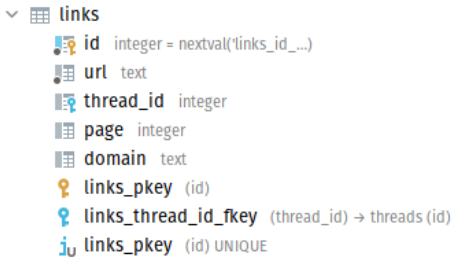

Figure 3: Links table schema

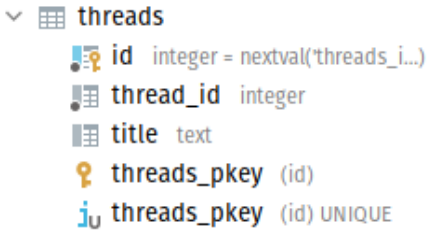

Figure 2: Threads table schema