

Graduate Project Proposal

Internet forums contain a wealth of information that I try to collect and apply to my own life. One facet of my life where I rely on such data is for product recommendations. In these days of online shopping, fake reviews, and third-party suppliers, it is becoming harder to sift through trustable product information on shopping websites. When looking for a certain product, I often search for what others have recommended on websites like forums Reddit and Bogleheads as well.

For this project, I plan on scraping through posts on **Bogleheads** (bogleheads.org) looking for **Amazon** links (amazon.com/...). I'll set up a database (local or online managed **PostgreSQL**) where the product link, the containing post, and thread information will be stored. For scraping and crawling through the forum, I have experience with using BeautifulSoup for scraping but would prefer to learn **Scrapy** since it can crawl through pages as well.

For data analysis and reporting, I would like to use **Pandas**. I found a tutorial online (<https://pythontic.com/pandas/serialization/postgresql>) for importing data into Pandas from a PostgreSQL database. Once in Pandas, I will use their official documentation (https://pandas.pydata.org/pandas-docs/stable/user_guide/) as a guide for manipulating and visualizing the collected data. I am hoping to display results including the last ten products recommended and all-time recommended products.

