Yasaman Sabersheikh

yasaman2@illinois.edu

CS410 – Fall 2021 – Text Transformation Systems

Final Project – 1 person team

Project Progress Report

Goodreads Scraper

SUMMARY

Goodreads is a free website on books that allows individuals to freely search for books and reviews. Users can sign up and generate library and reading lists. https://www.goodreads.com/

Finding books that are award winners are not easy on the website. Goodreads has a link for Reader's Choice awards but other than that it is hard to find other literary awards such as The Man Booker Prize or Giller Prize.
In this project I intend to write a web scraper to find the books that award winners. The result can be used for book recommendation, book clubs or a blog post.

PROJECT GOALS

1. Create a scraper to scrape the data in Goodreads website
2. Identify Award Winner books/authors

SPECIFICATIONS

- Python programming language for scripting
- Selenium to retrieve the Html source code
- BeautifulSoup to extract book-data from the Html source code

PROGRESS

- Researched what is on the web, best practices
- Learned Selenium and BeautifulSoup in detail
- Accessed “new releases” webpage through Google Chrome
- Using Selenium and Python Code to sign in to GoodReads website, by code, automatically.
- Grab all the “new releases” for last 12 months and categorized their information in Dictionary and DataFrame.

TASKS TO COMPLETE

- Go Deeper in the tree and scrape information inside the URL of each book, looking for the word Prize or Award
- 30% is complete

CHALLENGES

- As a one-person team, it was hard for me to start, as I could not brainstorm with anyone. I went through lots of materials to define my starting point.
- I started big, I wanted to find all the prize winner books on the website and all the best reviews. As the data is in different places on the website and I was dealing with massive amount of data I realized, it is not a 20h project if I consider both best reviewed and prize winners. I narrowed down the scope and decided to start small and add to it, as professor suggested. So, I started with prize winners of the year.
- When I was trying to scrape the html pages with BeautifulSoup, because of the structure of the pages and their html, I could not get the data I needed easily. It took me much longer than what I anticipated to get the data and load a Dictionary and DataFrame with it.