Finding Yesterday's Top 100 Books

This project demonstrates how to parse through HTML information by performing get requests to the Gutenberg website (www.gutenberg.com) using BeautifulSoup. It will find the page that displays Yesterday's Top 100 Books and parse through the information to print the list out nicely to the user. You will be able to see the steps of parsing cell-by-cell until the finish product is complete.

System Features

This scraper system will make get requests to the Gutenberg website, finding the yesterday's top 100 books page in order to parse through the HTML information. The data will be parsed in a loop step-by-step until it extracts the top 100 books list and prints it out for the user in an easy-to-read manner.

Installations and Requirements

This system will require the following Python libraries to be imported in order to retrieve information from the HTML, parse through the data, and handle secure connections:

- urllib.request
- requests
- re
- bs4 (BeautifulSoup)
- ssl

Using the System

You can use this web scraper system in either Jupyter Notebook or any other Python IDE, such as PyCharm. This system could also be run in a Python terminal. However, it is recommended to be used in an IDE, as that is where the script was created and run before. If you wish to use it in Jupyter Notebook, download the .ipynb file for use in your own Jupyter Notebook or copy each cell into your Jupyter Notebook. You may also copy and paste the code into another Python IDE if you prefer a different IDE besides Jupyter Notebook.

The script will make a get request to the Gutenberg website to retrieve yesterday's top 100 books. BeautifulSoup will parse through the HTML information cell-by-cell to clean the data. You will see the parsing and cleaning happening step-by-step within the script. Finally, it finishes with a beautiful, easy-to-read list of yesterday's top 100 books.

Contact

For any questions or concerns, please feel free to contact me, Ahria Dominguez, at ahriadominguez@outlook.com.