Web Scraping and Social Media Scraping

Project 2022

Participants:

Lavanya Ramaswamy, 441980

Ankit Pandey, 437949

Website:
https://www.goodreads.com/list/show/264.Books_That_Everyone_Should_Read_At_Least_Once

Description:

We have used the above website to perform scraping mechanism. The website is about all kind of books to read at least once. So, there are all the name of the books, the author's name, how much is the rating of each book and many more details. We have used scraping tools to scrape out information about the title of the book, the authors and numbered ratings. The scraping tools we have used are BeautifulSoup, scrapy and Selenium. We have used python idle and also chrome driver for the outputs. The outputs are shown in a excel sheet form i.e., csv file. This is the basic information regarding our project. Further, we will explain each and every scraping mechanism used.

The 3 scrapy mechanisms are as follows:

1. Beautiful Soup:
   BSoup is basically a python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping. First, you have to download the package in the cmd prompt using pip install. After getting installed you have to start the coding in python by importing bsoup first. Later you have to request URL by using requests. You have to provide the website you want to scrape. After getting the URL, you have to use the BeautifulSoup function paste your URL into it, which goes like this: soup = BeautifulSoup (source, 'lxml'). After this you just have to use basic codes for the information you want to scrape.

2. Scrapy:
   Scrapy is a free and open-source web-crawling framework written in Python. Originally designed for web scraping, it can also be used to extract data using APIs or as a general-purpose web crawler. Scrapy has to be downloaded and then similar to soup, it has to be coded in the python idle by starting with import scrapy. We have used scrapy crawler too. We have scraped information from the website into an csv file.

3. Selenium:
   Selenium is a open source umbrella project for a range of tools and libraries aimed at supporting browser automation. It also provides a test domain-specific language (Selenese) to write tests in Python and many more programming languages. Install the tool using pip installation and then import in python idle. Specify your website and then locate the things you want to scrape and extract the information into a python or csv file. A ChromeDriver is a separate executable or a standalone server that Selenium WebDriver uses to launch Google Chrome.

Description of participant work:

BeautifulSoup done by Lavanya Ramaswamy.

Selenium done by Ankit Pandey

Scrapy done by Ankit Pandey

Documentation and creating repository, uploading files in github, and submission done by Lavanya Ramaswamy.