CS-306
Data Analysis and Visualization
201701403-Megh Shah

Web Scraper

Abstract

Some websites can contain a very large amount of invaluable data.
If you wanted to access this information, you'd either have to use whatever format the website uses or copy-paste the information manually into a new document. Here's where web scraping can help.
Web scraping refers to the extraction of data from a website. This information is collected and then exported into a format that is more useful for the user. Be it a spreadsheet or an API.

Data displayed by most websites can only be viewed using a web browser. They do not offer the functionality to save a copy of this data for personal use. The only option then is to manually copy and paste the data - a very tedious job which can take many hours or sometimes days to complete. Web Scraping is the technique of automating this process so that instead of manually copying the data from websites, the Web Scraping software will perform the same task within a fraction of the time.

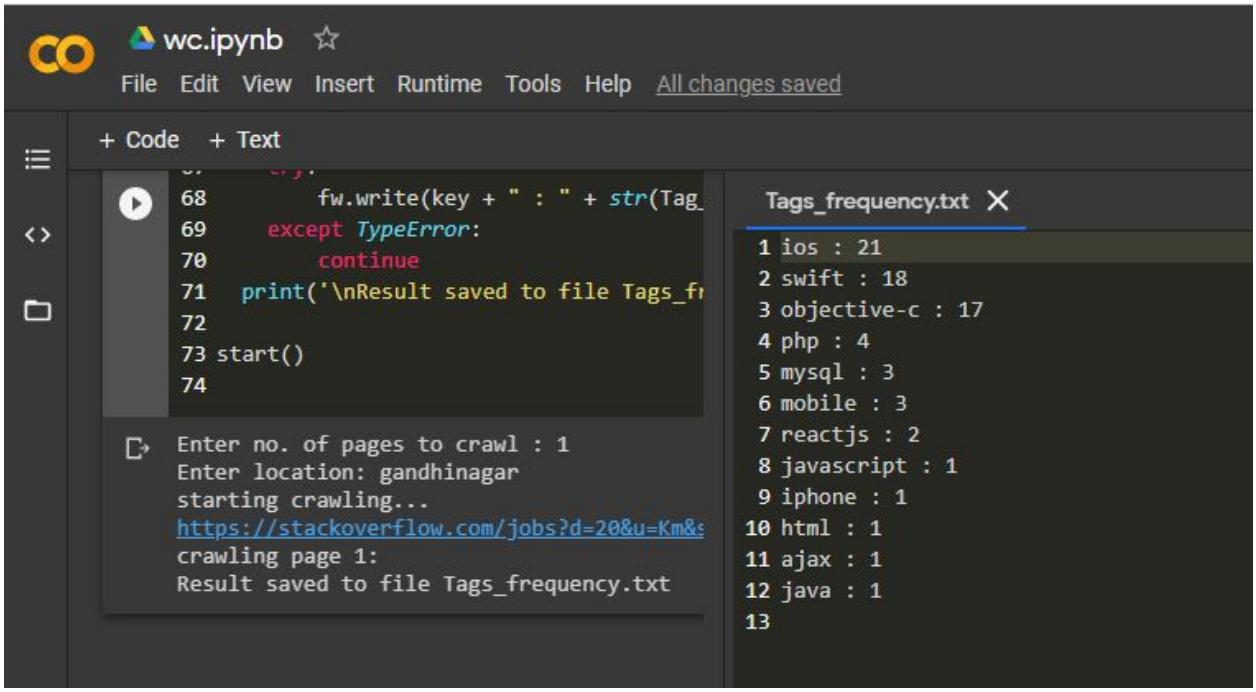
Here I have built a basic web scraper to get information from the **StackOverflow** website regarding the popularity of technologies based on the current job market.

Methodology:

- The number of pages to be crawled and the location of the job market to be checked is taken as input from the user.
- Based on this input a URL of the StackOverflow website is generated that contains the information regarding the available jobs in that location.
- Each job description has the required technical skills in the form of tags on the site(eg. Python, IOS, etc)

- The number of occurrences of each tag are counted for the number of pages entered.
- This is done by parsing the HTML file for the webpage using the BeautifulSoup library in Python
- Each is parsed from the HTML file with <class='ps-relative d-inline-block z-selected'> and the <a> tag is read from it.
- The result is stored in a text file(Tags_frequency.txt)

Outcome:



The text file displays the count of each technical skill in demand based on the currently available jobs thus giving an estimate about the popularity of the technologies in that location.

References:

https://www.webharvy.com/articles/what-is-web-scraping.html
https://www.parsehub.com/blog/what-is-web-scraping/
https://www.youtube.com/watch?v=XQgXKtPSzUI&t=732s
https://github.com/calebwin/frequent