WEBSCRAPYING PROJECT

http://dove.org/reviews

STUDENTS

Wojciech Misiura – 410579 and Muhammad Saqib Masood 437965

Participation of the members

Scrapy, Selenium and BeautifulSoup were created together by Muhammad Saqib Masood and Wojciech Misiura on google meets meetings and on-site.

Project Description

In this code we scrape documentary movies details from documentary reviews. To get data, we scraped http://dove.org/reviews site for documentary category. Dove.org is a site where Faith and Family-Focused Reviews for Today's Media reviews are posted.

Scraper Mechanism

BeautifulSoup

In BeautifulSoup, we use request library ("requests.get") to access website, re library for regular expressions and ssl to enable running BeautifulSoup scraper on windows. To access correct information, we find information by using soup.select:

- soup.select('div.content-info span:nth-child(1)')
- soup.select('div.content-info h4:nth-child(2)')
- soup.select('div.content-approved strong:nth-child(1)')

To scrape data from 100 pages, we implemented for loop, to get and store data we used numpy and pandas.

Scrapy

First excel file with 3 collumns is created. We initialize start_urls variable, define domain and list of pages which are to be scraped. To get informations, we use xpath, by providing class names. Then we create a dictionary from the lists, then dictionary to dataframe and values are returned to excel file.

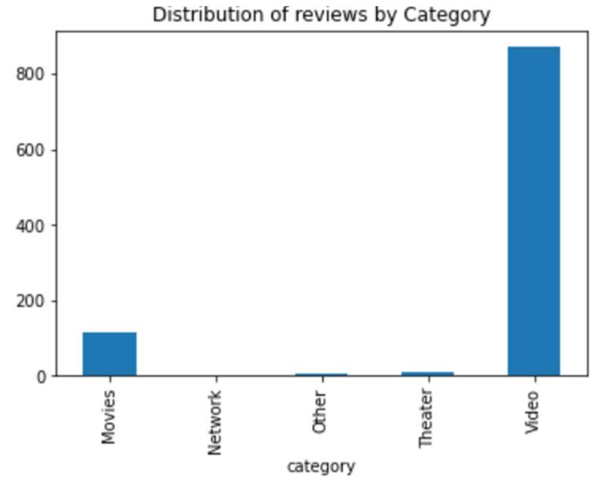
Selenium

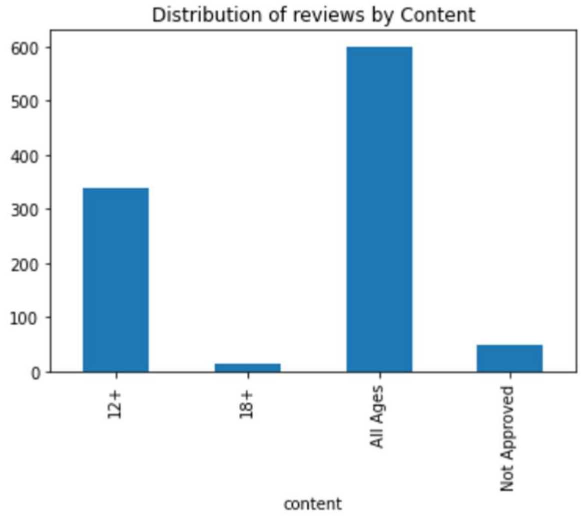
In Selenium scraper, we do not use links but use .click to get to the next page in the loop. This change is the main reason why Selenium is the slowest method we used. To get elements, we use xpath, returned values were saved into lists, after values are saved the button is clicked and scraper scrapes next site.

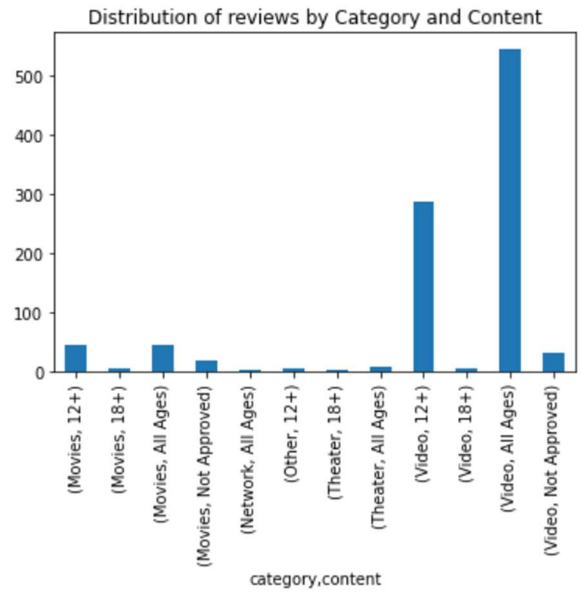
Technical Output

Output	**Description**
Category	Category of the movie
Title	Title of the movie
Content	Age restrictions

Graphical Analysis and comments







By looking at above graphs, we can see that there are few categories which are clearly underrepresented, such us categories “Network”, “Other” and “Theater”. Most documents reviewed are allowed for All Ages, with documentaries with Not Allowed age restriction outnumbering allowed only for age +18.

Performance analysis

Scraper	Time
Beautiful Soup 4	0:02:07:841878
Scrapy	0:00:11:533591
Selenium	0:05:08:895825

By looking at the table, it is clear that the fastest scraper is Scrapy, slowest is Selenium as described in – clicking buttons instead of using links.