**Scraping IMDB website using Python**

Made by-

* Vanshika Saini
* Shivangi Agarwal
* Varun Goel
* Vaibhav Garg

**Abstract**

We have scraped IMDB website to analyze the distributions of IMDB and Metacritic movie ratings and then store the scraped data in a CSV file.

Here we have used the get () function from the **Requests** module to request the server the content of the web page.

**Beautiful Soup** provides a few simple methods and Pythonic idioms for navigating, searching, and modifying a parse tree using Python parsers like lxml and html5lib. It automatically converts incoming documents to Unicode and outgoing documents to UTF-8.

Usually all the data we parse from the web page is not essential for the analysis. Therefore, it is required to first clean the data for extracting the useful information and for this we use **Pandas**. It is a python package which is widely used for the purpose of data munging and preparation. In our project we have used a function DataFrame (), which is a 2D data structure and is used to reshape the data in the form of a table. This makes the data readable.

Finally, we have used **Matplotlib** to plot the data in form of graphs and conclude our analysis.

**Introduction**

To source data for data science projects, you’ll often rely on [SQL](https://en.wikipedia.org/wiki/SQL) and [NoSQL](https://en.wikipedia.org/wiki/NoSQL) databases, [APIs](https://en.wikipedia.org/wiki/Application_programming_interface), or ready-made CSV data sets. The problem is that you can’t always find a data set on your topic, databases are not kept current and APIs are either expensive or have usage limits. If the data you’re looking for is on a web page, however, then the solution to all these problems is **web scraping**.

Through web scraping services unstructured data are converted into structured data which can be stored and verified in a centralized data bank. The aim is to collect, store and analyse data. The data analysis is very much needed in a society to extract any information and transforming it into a format helpful to interpret. Thus, web scraping services have a direct influence on the outcome which is needed from the data collection. Web data extraction is the process of transforming the useful content on websites into valuable business assets. There are several web extracting software that has emerged in the market which helps to address this problem. The software aids in extracting structured content from a web page and exposes the required services as APIs and makes it useable for further processing. It is necessary to know the available technologies in the market today. The available technologies that are related may be in different languages written such as java, python, php etc. The benefits of this are beyond the limitations of the users. Since there is rise in new online business through internet this has an adverse effect on the consumers as well. Online marketing analyst use web scraping methods to grab some information from other competitors such as emails, targeted keywords and links and also traffic source. The scraping techniques are used for personal as well as commercial usage. All the techniques available has its own pros and cons to overcome this there is need to have a clear idea on the usage of these techniques in social networking.

**Experimental Setup: -**

* Jupyter Notebook
* Requests
* BeautifulSoup
* Pandas
* Matplotlib

**Block Diagram**

Start

Copy IMDB URL

Request Data

Break loop

Requests Pg>72

Analyze Data

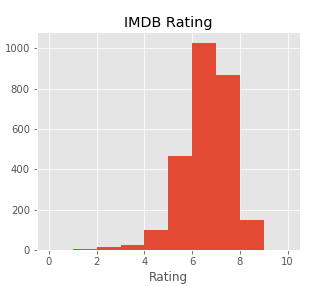
Data found

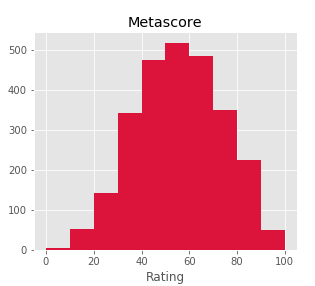
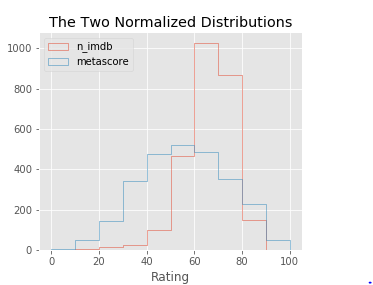
Scrape Info

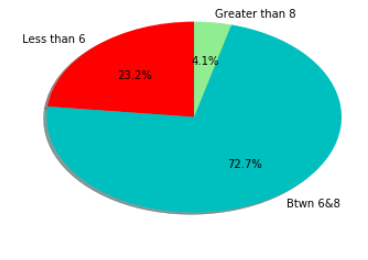
Plots graphs

Parsing data using BS

End

**Graphs**

****

****

**Conclusion**

* From the IMDB histogram, we can see that most ratings are between 6 and 8. There are few movies with a rating greater than 8, and even fewer with a rating smaller than 4. This indicates that both very good movies and very bad movies are rarer.
* The distribution of Metascore ratings resembles a normal distribution - most ratings are average, peaking at the value of approximately 50. According to this distribution, there are indeed fewer very good and very bad movies, but not that few as the IMDB ratings indicate.
* On the comparative graph, it’s clearer that the IMDB distribution is highly skewed toward the higher part of the average ratings, while the Metascore ratings seem to have a much more balanced distribution.
* One hypothesis for this result can be that many users tend to have a binary method of assessing movies. If they like the movie, they give it a 10. If they don’t like the movie, they give it a very small rating, or they don’t bother to rate the movie.

**Related Research**

1. Web Data Extraction, Applications and Techniques: A SurveyEmilio Ferraraa, Pasquale De Meob, Giacomo Fiumarac, Robert Baumgartnerd
2. WebSelF: A Web Scraping Framework Jakob Thomsen1, Erik Ernst1, Claus Brabrand2, and Michael Schwartzbach
3. Exploiting web scraping in a collaborative filtering- based approach to web advertising Eloisa Vargiu1, 2, Mirko Urru1
4. Dipartimento di Matematica e Informatica, Università di Cagliari, Italy. 2. Barcelona Digital Technology Centre, Spain
5. Faustina Johnson and Santosh Kumar Gupta.Web Content Mining Techniques: A Survey, International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012
6. Cohen and Fan. Learning page-independent heuristics for extracting data from web pages. CN, 31(11-16), 1999.

**References**: -

<https://www.crummy.com/software/BeautifulSoup/bs4/doc>

<http://pandas.pydata.org/pandas-docs/stable/10min.html>

<https://pythonprogramming.net>

<http://www.data-analysis-in-python.org/why_python.html>

<https://www.dataquest.io/blog/web-scraping-tutorial-python>

<https://www.dataquest.io/blog/web-scraping-beautifulsoup>