Scraping Static and Dynamic Webpages

Nirmal Kanagasabai

***Abstract*—Web scraping, as it is commonly called, has turned out to be a great technique to gather great volumes of data that is available online. Data enthusiasts have been employing different techniques to scrape the websites and make use of the content that is publicly available. In this exercise, the focus is on scraping two types of web pages - static and dynamic ones. A typical example of a static webpage is whosdatedwho.com which has a complete database of celebrity relationships. For the dynamic webpage module, the websites Empeopled and Boardest are chosen. Both Empeopled and Boardest are Reddit-esque sites which has infinite scrolling (where new posts are loaded dynamically when the user scrolls down to the bottom of the page). After scraping the content, a small analysis is carried out on the data and is reported. Also, the challenges and the ways to overcome it that one runs into while finding the ideal scraper will be discussed.**

I. Introduction

WEB scraping has become quite inevitable for data enthusiasts considering the exabytes of data that are offered by the internet and social media. Most importantly, web scraping are used by research and social media companies. It is also used by website creators to scrape the content of a website they are interested in. The most common applications include scraping of online shopping websites, social media sites, meteorological websites, travel recommendation sites, job forums, etc. The list of use cases is endless.

The common saying across all web scrapers is, "Any content that can be viewed on a webpage can be scraped. Period". However, there is no fixed structure to extract all the content. This is mainly because, each site is organized in a different way. Few of them are static and few of them, dynamic. These circumstances make web scraping an integral component of data scientists.

II. Scraping Static Webpages

The static web pages, as the name states, are 'static'. The entire web content is available on the first load. This makes it easier for the data enthusiasts to identify what they need and define a way to pick only them from the internet. Despite the availability of website-copiers like WGET and HTTrack, the data engineers work on retrieving data only from specific fields or elements in the website. This, not only saves time but also a lot of space which the entire website will occupy. Also, doing this will not call for data processing once the content is scraped. Whosdatedwho.com had a big database of celebrity relationships (including the rumoured ones). Scraping all the relationships from this site with the start and end dates was a challenge. It involved two steps:

Advisor: Prof. Derek Ruths, McGill University, 2017.

Step 1: To identify the list of celebrities with their names starting with a particular alphabet. This was done to ensure that no celebrity was left behind during the scraping process. The base_url that was used to scrape the content was "http://www.whosdatedwho.com/popular?letter='alphabet'" where the letter from 'a' to 'z' was substituted in the place of 'alphabet' to complete the query. This page had 'infinite scrolling' as the list gets populated as and when the user scrolls down. Ideally, this was yet another exercise of scraping a dynamic webpage. To achieve this, Selenium web driver was made use of to automatically control the browser and scroll down until the last element loads. As and when the new element loads, the hyperlinks of different users are scraped and is stored in a .csv file.

Step 2: The second step was relatively easier. BeautifulSoup, a Python library for screen scraping was used. The list of urls is fetched from the .csv file and one after another (http://www.whosdatedwho.com/dating/'celebrity-name'), each one was used to query and BeautifulSoup scrapes the corresponding elements that is needed and saves them in another .csv file. As per the requirement, for each celebrity, their name, the list of partners they were in a relationship with, their type of relationship and the start and end dates of the relationship was scraped. For the celebrity name, the 'topic' tag was used and for the other entities, a 'div' tag with an id attribute of 'ff-dating-history-table' was made use of to find the table and the entries.

Challenges: The biggest challenge in this exercise was to find a way to gather all the URLs that corresponds to celebrity profiles. The 'list' that they had in the website didn't have all the celebrities in it. However, the 'popular' category covered all the users in the forum. Also, these two pages, being infinite scrolling webpages, gave a challenge while extracting the content using BeautifulSoup. Thankfully, Selenium Web-drivers comes to the rescue. Through this, the browser is automated, the webpage is loaded and is scrolled until the end of the page is reached. Also, with the sophisticated functions offered by selenium web drivers, the corresponding elements could be identified by their XPaths and can be retrieved.

III. Scraping Dynamic Webpages

This exercise was similar to the step 1 of the previous section. The webpages that were expected to be scraped were infinite scrolling and was rendered using JavaScript. Due to this, using a simple static scraper is ruled out. Yet again, Selenium Web Driver was made use of to automate the browser and scroll to the bottom of the page. The expectation was to collect the latest 1000 posts from the Reddit-esque

site Empeopled.

However, there was a limitation in this case. Empeopled displayed only the latest 300 posts in their public 'Home' page. Using the find_elements function of Selenium web driver, the following elements were scraped from the webpage - title, username, vote_count, date and the tags. As this doesn't go well with the requirement, another dynamic website (Boardest) was considered. This webpage was also an infinite scrolling page and is rendered using JavaScript. The same technique was employed to scrape title, username, story (description), date and the tags. Looking at the site, it was obvious that there wasn't any upvotes or comments in the posts. Hence, scraping those elements would not serve any purpose.

IV. Analysis

Apart from scraping the contents from all these sites, small analysis was carried out with the dataset at hand. For the celebrities dataset scraped from Whosdatedwho.com, the celebrities with maximum number of relationships was calculated and has been listed down in Table I. Looking deeper into the social profiles of all these celebrities, most of their relationship types were 'encounters' or 'short-term-relationships'. Out of this list, all of them are American actress or actress (predominantly in the 20th century). The top of the list 'Rocco Siffredi' was a porn star and his relationship types have mostly been 'encounters'. It is also surprising to find that he is the only married person in this list. The rest are all single which also proves that they don't stick to the same partner for a long time (with a lot of short-term relationships). Also, a graph was plotted with the count of different relationship types. Fig. 1

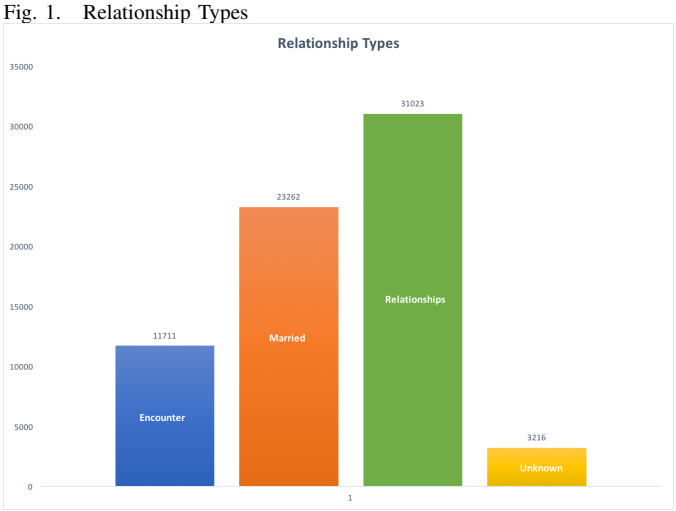
In the dataset extracted from Empeopled and Boardest webpages, the tags to which the posts are associated with are compared. Initially, the number of posts per particular tag was calculated. This was followed by sorting the list in descending order to find which tags contribute the most. It can be observed that the users are obsessed with few topics over another. In both the figures Fig. 2 and Fig. 3, it can be observed that, the tags which contribute the most 'News', 'Science' and 'Technology' are repeated on both the forums. They, along with few more, contribute to more than 50% of the posts leaving the rest of the tags to take up only 31% and 34% respectively.

To understand the webpages better, a small comparison was made on the commonly repeated tags namely - 'Science and Technology', 'News', 'Travel', 'Health and Lifestyle', 'Arts and Entertainment', 'Animals', 'Music', 'Sports' and 'Gaming' as shown in Fig. 1. Though the datasets are skewed where one has 1000 posts and the other has 300 posts, with the tag% comparison that was carried out, it can be observed that, Boardest has more tags compared to Empeopled. Few tags like 'Science and Technology' are split into two - 'Science' as well as 'Technology'. Another possible inference from the comparison is that the users of Empeopled are diverse and discuss a lot about 'news', 'politics', 'general' and 'Science and Technology' compared to the users of Boardest who discuss a lot about technologies. For example, the tags that have a lot of articles include, 'Technology', 'Internet', 'Science', 'Programming', 'Computers', 'Mobile Devices', 'Graphic Design', 'Webmsters', 'Bitcoin and Cryptocurrency', etc. With this, we can conclude that Boardest proves to be an ideal platform for Tech-savvy people compared to the diverse nature of Empeopled where all sectors of people are included.

TABLE I
Top 10 Celebrities with maximum number of relationships

Celebrity	No. of Relationships
Rocco Siffredi	124
Warren Beatty	105
Joan Crawford	89
Lana Turner	83
Karrine Steffans	78
Lindsay Lohan	76
Tallulah Bankhead	69
Ava Gardner	67
Clark Gable	56
Paris Hilton	53

Fig. 1. Relationship Types



V. Conclusion

The above exercises prove to be ideal for beginners who wishes to learn about different types of webpages and how to scrape content from them. It also introduces them to a lot of new frameworks and libraries which can be made use of in the process. The challenges that was discussed above was encountered and tackled after many attempts. Hence, this report will act as a guide offering first hand information on how to scrape websites efficiently.

Fig. 2. Tags in Boardest

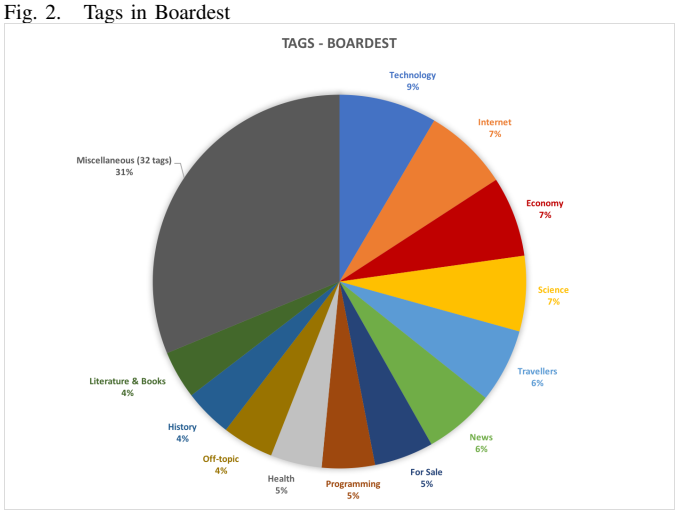


Fig. 3. Tags in Empeopled

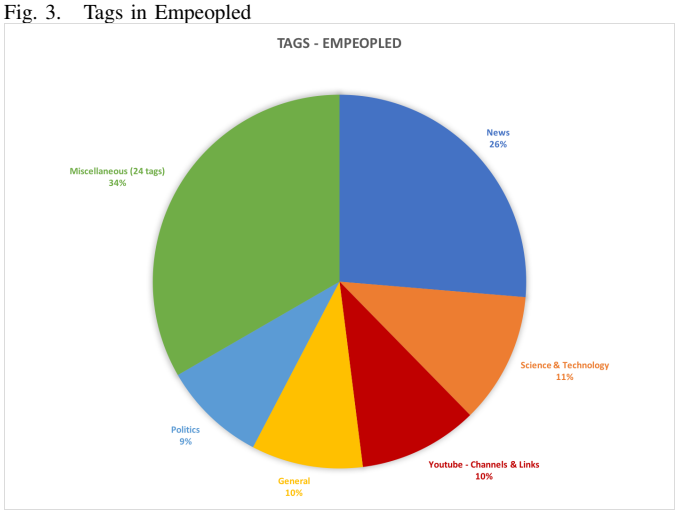


Fig. 4. Tag Comparison: Empeopled vs. Boardest