Data Scraping and Data Representation Case Study

Pharmascroll Research and Consulting Private Limited

The Problem:

As a research company we have to prepare report on several datasets and for that we need a data scraping tool and a platform where we can represent the data. In this case study we will simply scrape data from a website about drugs and represent it on a web-based platform.

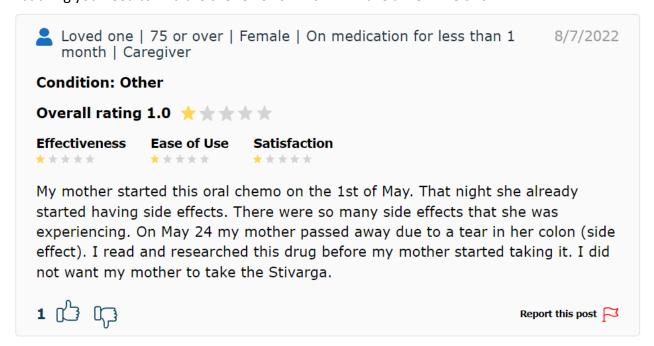
What you have to do:

Part 1: Data Scraping and Saving

You will be provided with a csv file of multiple links from the same website which contains the information about some drugs which we need to research on.

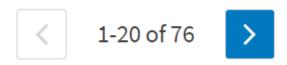
The URL be like https://www.webmd.com/drugs/drugreview-162543-stivarga-oral?drugid=162543&drugname=stivarga.

First thing you need to find are the reviews which will have a view like this:

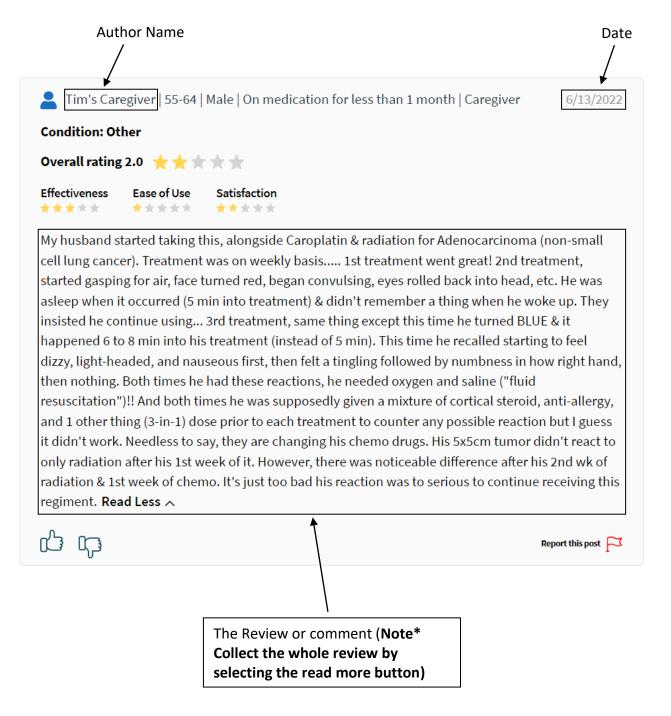


The reviews will contain a lot of data and as well as there will be multiple reviews for a drug on the website. So, you need to extract the data mentioned below from all of the reviews.

As well as there can be a lot of reviews which will be not present on the same page so you have to go to the next page and collect all of the reviews as well. The pagination for next page will look like this:



The data you need to extract from a review



Data to be collected:

- 1. Author name
- 2. Date of review
- 3. Comment

After you collect all the above from a review you must either store it in a database or a csv file. Database can be any (SQL, MongoDB etc).

Your database or csv should have these headers:

Date, Author Name, Content, Keyword, URL

Part 2: Data Representation

After you have successfully collected all the reviews from the website and stored them in the database you have to represent the data on a web-based portal. If you do not have an hosting and database you can create the portal on localhost as well.

Note*- There is no language barrier you can use any scripting and web development language.

How to represent data:

Views can be accordingly your creativity, but you have to represent it in this format-

Review 1		Date
Keyword or Drug Name		URL
Review Content		
Review 2		Date
Keyword or Drug Name		URL
Review Content		

Note*- You must share the source code for the data scraping and web portal as well.