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

Author Name   Date

Tim's Caregiver | 55-64 | Male | On medication for less than 1 month | Caregiver   6/13/2022

Condition: Other

Overall rating 2.0 ★★☆☆☆

Effectiveness   **Ease of Use**   **Satisfaction**
★★★☆☆   ★★★★★   ★★★★★

My husband started taking this, alongside Caroplatin & radiation for Adenocarcinoma (non-small cell lung cancer). Treatment was on weekly basis..... 1st treatment went great! 2nd treatment, started gasping for air, face turned red, began convulsing, eyes rolled back into head, etc. He was asleep when it occurred (5 min into treatment) & didn't remember a thing when he woke up. They insisted he continue using... 3rd treatment, same thing except this time he turned BLUE & it happened 6 to 8 min into his treatment (instead of 5 min). This time he recalled starting to feel dizzy, light-headed, and nauseous first, then felt a tingling followed by numbness in how right hand, then nothing. Both times he had these reactions, he needed oxygen and saline ("fluid resuscitation")!! And both times he was supposedly given a mixture of cortical steroid, anti-allergy, and 1 other thing (3-in-1) dose prior to each treatment to counter any possible reaction but I guess it didn't work. Needless to say, they are changing his chemo drugs. His 5x5cm tumor didn't react to only radiation after his 1st week of it. However, there was noticeable difference after his 2nd wk of radiation & 1st week of chemo. It's just too bad his reaction was to serious to continue receiving this regiment. **Read Less** ∧

👍 👎   Report this post 🚩

The Review or comment (**Note* Collect the whole review by selecting the read more button**)

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.

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

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.