Applied Data Science

COMS30050/COMS30051/COMSM0055/COMSM0056 (2020 TB-2)

Lab 1: Data Scrapping from the Web

Week #1 Lab builds on the lectures on Data Ingress and involves scrapping data from the Web. There are two sets of tasks in this lab.

The first set of tasks involve using Beautiful Soup package to extract data from Guardian newspaper's website. These tasks build on the code in iPython notebook *05_web_scraping_beautiful_soup.ipynb.*

The second set of tasks in this lab involve using the Guardian's newspaper's REST API to extract news data. These tasks build on the code in iPython notebook *06_web_scraping_api-initial.ipynb*.

Using Beautiful Soup

Use Beautiful Soup to complete the following tasks:

Task 1. Extracting Linked News Stories

URL: https://www.theguardian.com/world/2021/jan/21/johnson-raises-fears-of-covid-lockdown-in-england-continuing-into-summertime

Open the above URL in your browser. You will notice that parts of the main news story are hyperlinked to other news stories published previously. For instance, the third paragraph is linked to a new story on Boris Johnson's visit to flood-hit Manchester. Your first task is to extract the links to these other news stories in the main news.

Task 2. Extracting Topics or Categories

URL: https://www.theguardian.com/world/2021/jan/21/johnson-raises-fears-of-covid-lockdown-in-england-continuing-into-summertime

Guardian's website tags the news story with a list of topics. Your second task is to find these topics.

Task 3. Listing All News Stories in a Section

URL: https://www.theguardian.com/uk/technology

In the chaining code block in *05_web_scraping_beautiful_soup.ipynb*, we tried listing all news stories on the technology page of the Guardian's website. You will notice that "js-headline-text" class fetches non-technology stories as well.

Your task is to filter only technology related stories that are listed under the Technology division of the Guardian's webpage

Task 4. List 50 Most Recent Technology-Related News Stories

URL: https://www.theguardian.com/uk/technology

On Guardian's technology home page, you will notice a link to “All Stories.” If you cannot locate it visually, use the browser’s find tool (Ctrl + F) and search for “All Stories.” Click the link to “All Stories” and observe the structure of the web page listing all stories. Here is the direct link to that web page: https://www.theguardian.com/technology/all

Your task is to extract 50 most recent technology stories published by Guardian.

Bonus Task. List All Technology-Related New Stories Published After January 25, 2021

URL: https://www.theguardian.com/technology/all

Using Guardian’s REST API

Use Guardian's API to complete the following tasks:

Task 6. Response Statistics

Use Guardian's API to identify the count of all news stories published under the Technology section. List the page size and the number of pages these results are displayed.

Task 7. News Stories About a Specific Topic

Return all stories in the technology section that are about privacy. Filter the stories that talk about WhatsApp and Signal.

Are there any privacy stories talking about privacy, WhatsApp, and Signal that do not talk about AI? List these stories.

Note. People write AI as AI or artificial intelligence.

Other search queries to try:

a. *All News Stories About a Phrase:* Return all news stories that are about stock squeeze. List the ones that are in the business section of the Guardian.
b. *All News Stories About a Person:* Return all news stories about Elon Musk published by Guardian in 2020 and 2021. How many of these news stories are about bitcoin? Of the stories that are about Elon Musk and Bitcoin, how many of those do not mention Tesla?

Task 8. Requesting Specific Content Using the API

Fetch the *i*th result from the list obtained from on the privacy related search query formed in Task 7. Identify the id of the ith result and fetch the headline and body text of the news story.

Hint 1: Obtain the content id using the following code

```
i = 0
api_url = response['results'][i]['apiUrl']
api_id = response['results'][i]['id']
```

Hint 2: Use the content id (attribute: ids) to fetch specific content

```
base_url = "https://content.guardianapis.com/search?"

search_string = "ids=%s&api-key=%s&show-fields=headline,body" %(api_id,
myapikey)

url = base_url + search_string
```

Task 9. Simple Text Processing to Prepare Data for Analysis

Print the body text of the news story. You will notice the body text contains some HTML tags. Clean the body text using string operations.

Using the clean body text: (1) Count the number of words and the number of unique words in the news story; and (2) Count occurrence of each word and store in a Pandas data frame

You could also apply regular expressions to clean the body text and compute the counts and word occurrences.