Sahil Khanna
U86445364

# Assignment 1
# Web Analytics and Mining

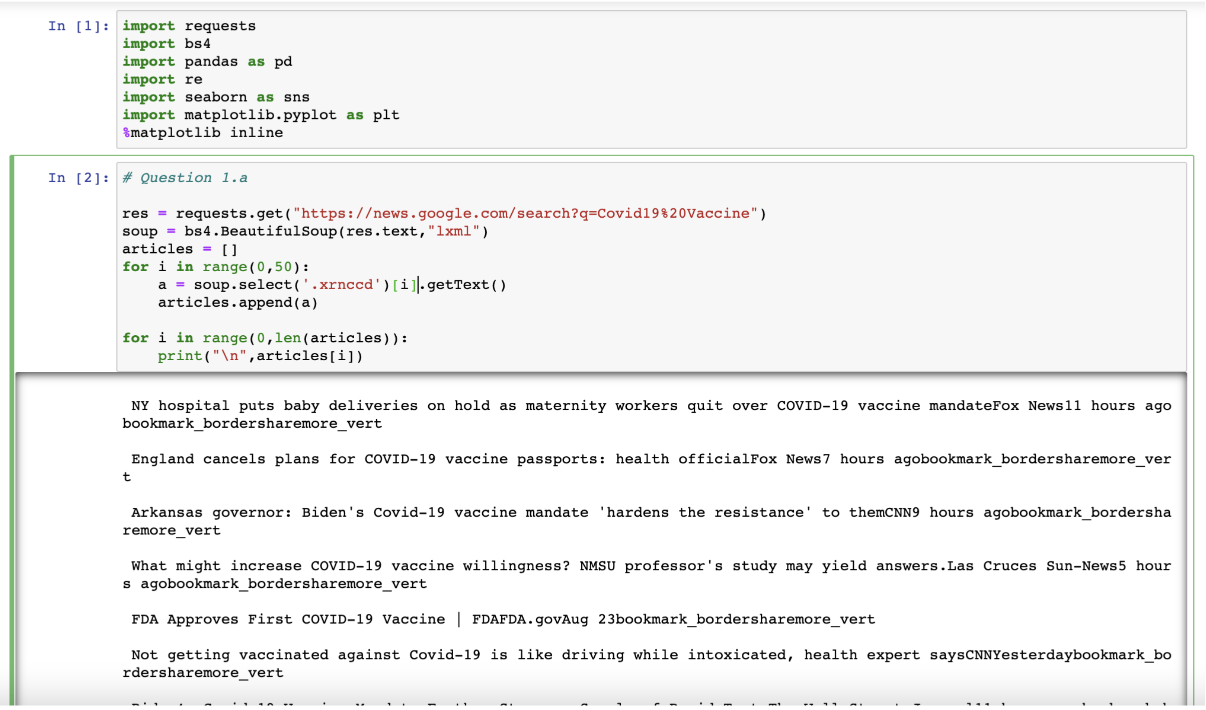
MET CS 688

The objective of this homework is to collect a list of articles in Google News or Bing News, which include the keyword Covid-19.

**(1)** Do either one of the following:

**(a)** Use a BeautifulSoap, RVEST, or any other libraries you like and search for the "Covid19 Vaccine" and download 50 articles title or abstract, from the search result, via web scrapping.

**Output:**



```
In [1]: import requests
        import bs4
        import pandas as pd
        import re
        import seaborn as sns
        import matplotlib.pyplot as plt
        %matplotlib inline
```

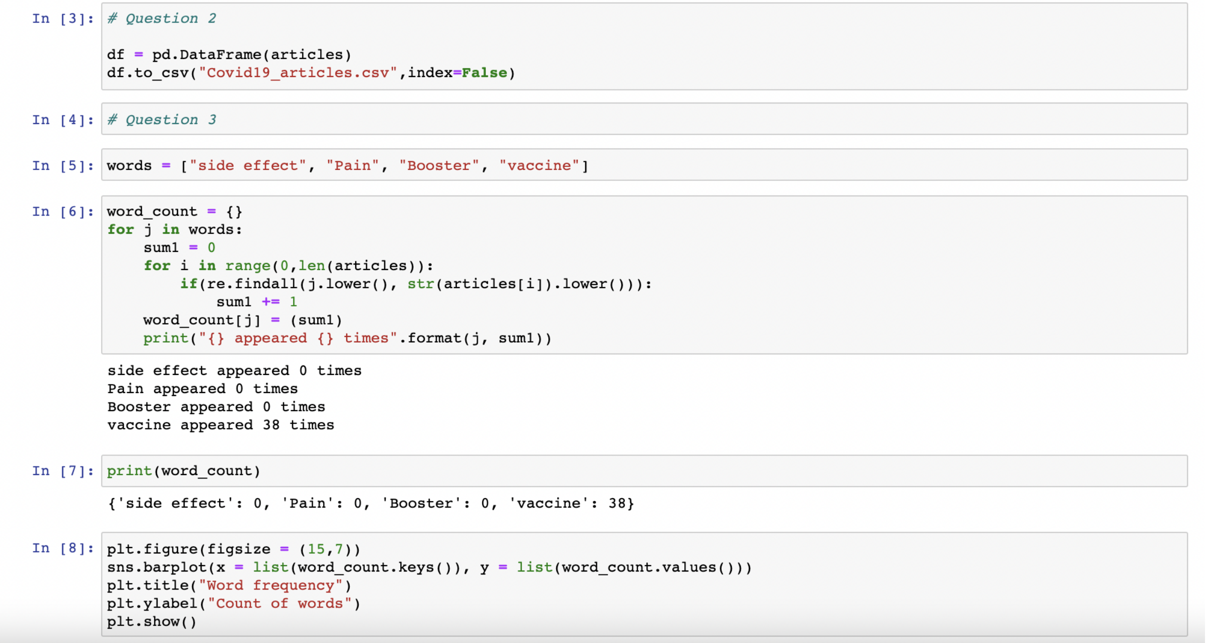
```
In [2]: # Question 1.a

        res = requests.get("https://news.google.com/search?q=Covid19%20Vaccine")
        soup = bs4.BeautifulSoup(res.text,"lxml")
        articles = []
        for i in range(0,50):
            a = soup.select('.xrnccd')[i].getText()
            articles.append(a)

        for i in range(0,len(articles)):
            print("\n",articles[i])
```

```
 NY hospital puts baby deliveries on hold as maternity workers quit over COVID-19 vaccine mandateFox News11 hours ago
bookmark_bordersharemore_vert

 England cancels plans for COVID-19 vaccine passports: health officialFox News7 hours agobookmark_bordersharemore_ver
t

 Arkansas governor: Biden's Covid-19 vaccine mandate 'hardens the resistance' to themCNN9 hours agobookmark_bordersha
remore_vert

 What might increase COVID-19 vaccine willingness? NMSU professor's study may yield answers.Las Cruces Sun-News5 hour
s agobookmark_bordersharemore_vert

 FDA Approves First COVID-19 Vaccine | FDAFDA.govAug 23bookmark_bordersharemore_vert

 Not getting vaccinated against Covid-19 is like driving while intoxicated, health expert saysCNNYesterdaybookmark_bo
rdersharemore_vert
```

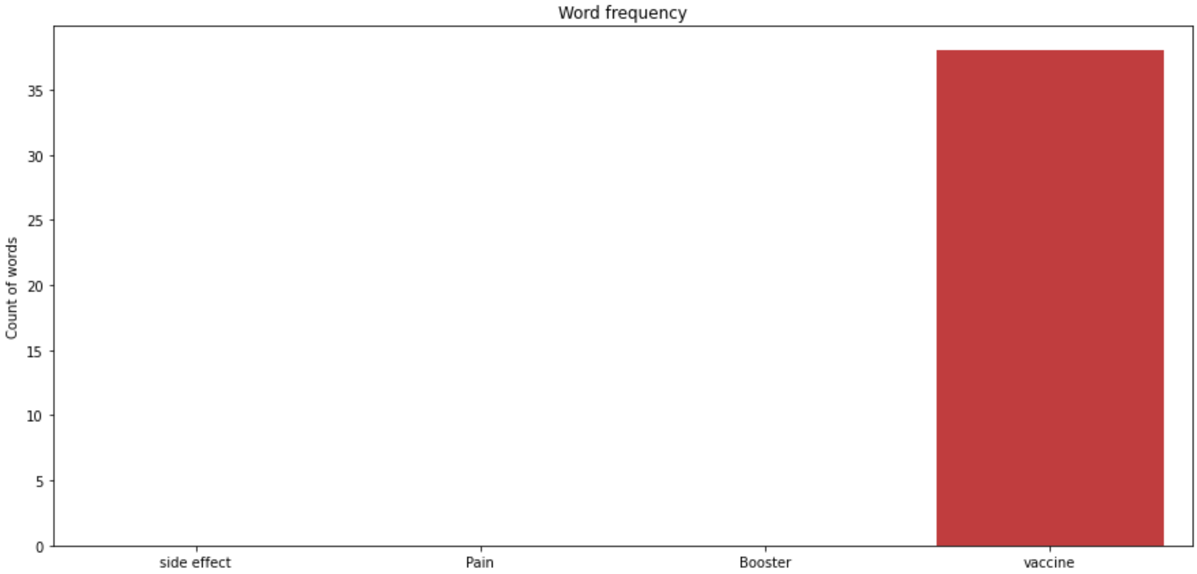
**(2)** Stores the downloaded articles' title and abstract on your local drive in a standard file format, e.g., CSV, XML, SQL, or JSON.

**(3)** Write a script to look at the title of each article and count the following words: "side effect," "Pain," "Booster," "vaccine" Then, create a word frequency histogram.

**Output:**



```
In [3]: # Question 2

        df = pd.DataFrame(articles)
        df.to_csv("Covid19_articles.csv",index=False)

In [4]: # Question 3

In [5]: words = ["side effect", "Pain", "Booster", "vaccine"]

In [6]: word_count = {}
        for j in words:
            sum1 = 0
            for i in range(0,len(articles)):
                if(re.findall(j.lower(), str(articles[i]).lower())):
                    sum1 += 1
            word_count[j] = (sum1)
            print("{} appeared {} times".format(j, sum1))

        side effect appeared 0 times
        Pain appeared 0 times
        Booster appeared 0 times
        vaccine appeared 38 times

In [7]: print(word_count)

        {'side effect': 0, 'Pain': 0, 'Booster': 0, 'vaccine': 38}

In [8]: plt.figure(figsize = (15,7))
        sns.barplot(x = list(word_count.keys()), y = list(word_count.values()))
        plt.title("Word frequency")
        plt.ylabel("Count of words")
        plt.show()
```


Word frequency
Count of words
side effect
Pain
Booster
vaccine