Assignment I – IMT 547

Due	Feb 1st by 5pm (See late policy on Canvas)
What to submit on Canvas	• Jupyter notebook (pdf and/or html) with answers to the questions listed below. ***Use proper code formatting. Use markdown cells to write questions and descriptive answers. Use code cells to insert your code, run your code and show output.*** • .csv data files: `pandemictweets.csv`, `WHOandUNtweets.csv`, `pandemicReddit.csv`

In the last several months social media platforms have been inundated with information related to the pandemic. Your task in this assignment is to obtain data from at least two social media platforms capturing some information about the pandemic.

Q1. Collecting Twitter data

1a). You plan to use Twitter's search API to collect data related to the coronavirus pandemic. Check Twitter to make a list of keywords that you would use to start this data collection. Provide your reason behind your choice. Your list should have at least 3 keywords.

1b). **Write code** to collect tweets related to the coronavirus pandemic. You can pick any combination of keywords that you selected in 1a). Save tweets in .csv file named: `pandemictweets.csv`. Your file needs to have at least 1000 rows of data.

Q2. Collecting Twitter user data

The World Health Organization (WHO) and the United Nations (UN) have both been very active in sharing COVID related information on Twitter. Here are their official Twitter handles: https://twitter.com/WHO, https://twitter.com/UN

2a). Using twitter API and `tweepy Cursor`, **write code** to obtain at least 5 pages of tweets from each of these two user accounts.

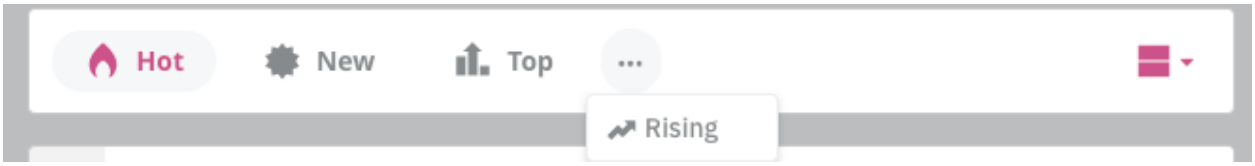
From the resulting data, parse at least 5 fields (example fields: `location, friends_count, created_at`). Now create a pandas dataframe, where you will have all this information. Print and show the `head()` of the dataframe to display the first few rows of data.

2b). **Write code** to turn the dataframe into a CSV and save it in a file: '`WHOandUNtweets.csv`'.

Q3. Reddit data collection

3a). You plan to use Reddit to find communities (subreddits) that have been actively discussing the pandemic. Check Reddit to make a list of communities that you would use to start this data collection. Your list should include at least 3 subreddits. Provide your reason behind your choice.
Hint: For example, r/Coronavirus is one such subreddit. There are more.

3b). Using the Reddit API you can fetch subreddit submissions which are *hot, new, rising, top.*
You decided to first fetch the 'rising' submissions.



For the list of subreddits that you just compiled in 3(a), for each of those, fetch at least 25 rising submissions.
For each submission, extract at least 10 fields.
Now create a `pandas dataframe`, where you will have all this information, subreddit name, the submission from each of these subreddits along with all the fields.

Print the `head()` of the dataframe to show the first few rows of data. **Write code** to answer 3(b).

3c). Turn the dataframe into a CSV file and save it in a file: '`pandemicReddit.csv`'.

3d). For each of the subreddits, you also decided to fetch the top 50 'hot' submissions and see which one of them (if any) overlaps with the top 50 'rising' submissions. **Write code** to find out.

BONUS question

For each of the subreddits that you picked in 3a), write code to find the rules for each subreddit. Print the short_name of each rule and the timestamp.
Hint: For example, for the subreddit("news"), on printing the rules and timestamp, example output should look like the following:

```
Not news 2016-01-26 06:24:11
Opinion/analysis or advocacy piece 2016-01-26 06:27:59
Politics 2016-01-26 06:31:33
Title not from article/editorialized title 2016-01-26 06:35:51
Paywall or is blogspam/steals content 2016-01-26 06:40:33
Covers an already-submitted story 2016-01-26 06:44:40
Racist, sexist, vitriolic, or overly crude 2016-01-26 06:47:09
Unnecessarily rude or provocative 2016-01-26 06:49:35
Cheap or distracting joke or meme 2016-01-26 06:51:12
Breaks sitewide rules, witchhunting 2016-01-26 06:56:47
```

Write code to answer the bonus question