Homework Assignment #2

Analysis of Twitter JSON data using Python

Syracuse University
Prof. Landowski
Student: Patrick Walsh
Date: 5/20/2023

The data and its source

The 2020–2021 Indian farmers' protest was a protest against three farm acts that were passed by the Parliament of India in September 2020. Farmer unions and their representatives demanded that the laws be repealed and stated that they would not accept a compromise.

In light of this protest, social media users voiced their opinion about the matter using the hashtag #FarmersProtest". The dataset consists of a JSON file, containing raw data about the tweets that have the hashtag "#FarmersProtest" (tweeted during February 2021).

The dataset can be found here:
https://www.kaggle.com/datasets/prathamsharma123/farmers-protest-tweets-dataset-raw-json

The dataset did not need much preprocessing aside from reading in the JSON into a Pandas dataframe. I checked for null values and filled any null values with a zero.

```
# Read JSON file containing tweets data and removce tweets not in English
file = 'farmers-protest-tweets-2021-2-4.json'
raw_tweets = pd.read_json(file, lines=True)
raw_tweets = raw_tweets[raw_tweets['lang']=='en']
print("Shape: ", raw_tweets.shape)
raw_tweets.head(5)
```

```
Shape:  (48429, 21)
```

	url	date	content	renderedContent	id	
0	https://twitter.com/ArjunSinghPanam/status/136...	2021-02-24 09:23:35+00:00	The world progresses while the Indian police a...	The world progresses while the Indian police a...	1364506249291784198	{'u 'ArjunSin 'displ
1	https://twitter.com/PrdeepNain/status/13645062...	2021-02-24 09:23:32+00:00	#FarmersProtest \n#ModiIgnoringFarmersDeaths \...	#FarmersProtest \n#ModiIgnoringFarmersDeaths \...	1364506237451313155	{'u 'Pr 'dis
3	https://twitter.com/anmoldhaliwal/status/13645...	2021-02-24 09:23:16+00:00	@ReallySwara @rohini_sgh watch full video here...	@ReallySwara @rohini_sgh watch full video here...	1364506167226032128	{'u 'anm

A brief description of the program

Attached to this report is a Jupyter Notebook program written in Python. The program requires the file ‘farmers-protest-tweets-2021-2-4.json’ to be read into a Pandas dataframe to run.

The program does data processing and analysis to answer the set of questions contained within this report and which are stated below.

Question 1

Bin the Tweets by day and report on the number of tweets per day.

1. Which day had the most Tweets?
2. Which day had the least number of Tweets?
3. What was the general trend of Tweets overtime?

To answer the first question, we first need to verify the month(s) in which these Tweets occurred. They should all be in February, but we can confirm that here:

```
# check for all the months represented
raw_tweets.date.dt.strftime('%m').unique()
```

```
array(['02'], dtype=object)
```

output above confirms that all the Tweets occurred in February.

We can then check which days these Tweets occurred on and add the month and day each Tweet occurred on to the dataframe:

```
# Now check which days these Tweets occurred on.
raw_tweets.date.dt.strftime('%d').unique()
```

```
array(['24', '23', '22', '21', '20', '19', '18', '17', '16', '15', '14',
       '13', '12'], dtype=object)
```

Add a column called 'date_bin' which captures the month and day the Tweet occurred on.

```
raw_tweets['date_bin'] = raw_tweets.date.dt.strftime('%m/%d')
```

Now let's check how many Tweets were sent on each day:

```
1 raw_tweets['date_bin'].value_counts()
```

```
02/17    4516
02/12    4423
02/15    4407
02/14    4350
02/16    4343
02/13    4124
02/20    3730
02/18    3701
02/23    3566
02/19    3540
02/22    3133
02/21    3120
02/24    1476
Name: date_bin, dtype: int64
```

Using the code below, we can answer the question, which is **February 17, which had 4,516 Tweets.**

1. Which day had the most Tweets?

```
1 # Which day had the most Tweets?
2
3 most_tweets = raw_tweets['date_bin'].value_counts()[0]
4 highest_day = raw_tweets['date_bin'].value_counts().idxmax()
5
6 print('The day with the highest number of Tweets was:', highest_day)
7 print('Number of Tweets on this day:',most_tweets)
```

```
The day with the highest number of Tweets was: 02/17
Number of Tweets on this day: 4516
```

2. Which day had the least number of Tweets?

```
# Which day had the least number of Tweets?

least_tweets = raw_tweets['date_bin'].value_counts()[-1]
lowest_day = raw_tweets['date_bin'].value_counts().idxmin()

print('The day with the lowest number of Tweets was:', lowest_day)
print('Number of Tweets on this day:',least_tweets)
```

```
The day with the lowest number of Tweets was: 02/24
Number of Tweets on this day: 1476
```

Finding the day with the least number of Tweets is shown above, the answer being **February 24, with 1,476 Tweets.**

3. What was the general trend of Tweets overtime?

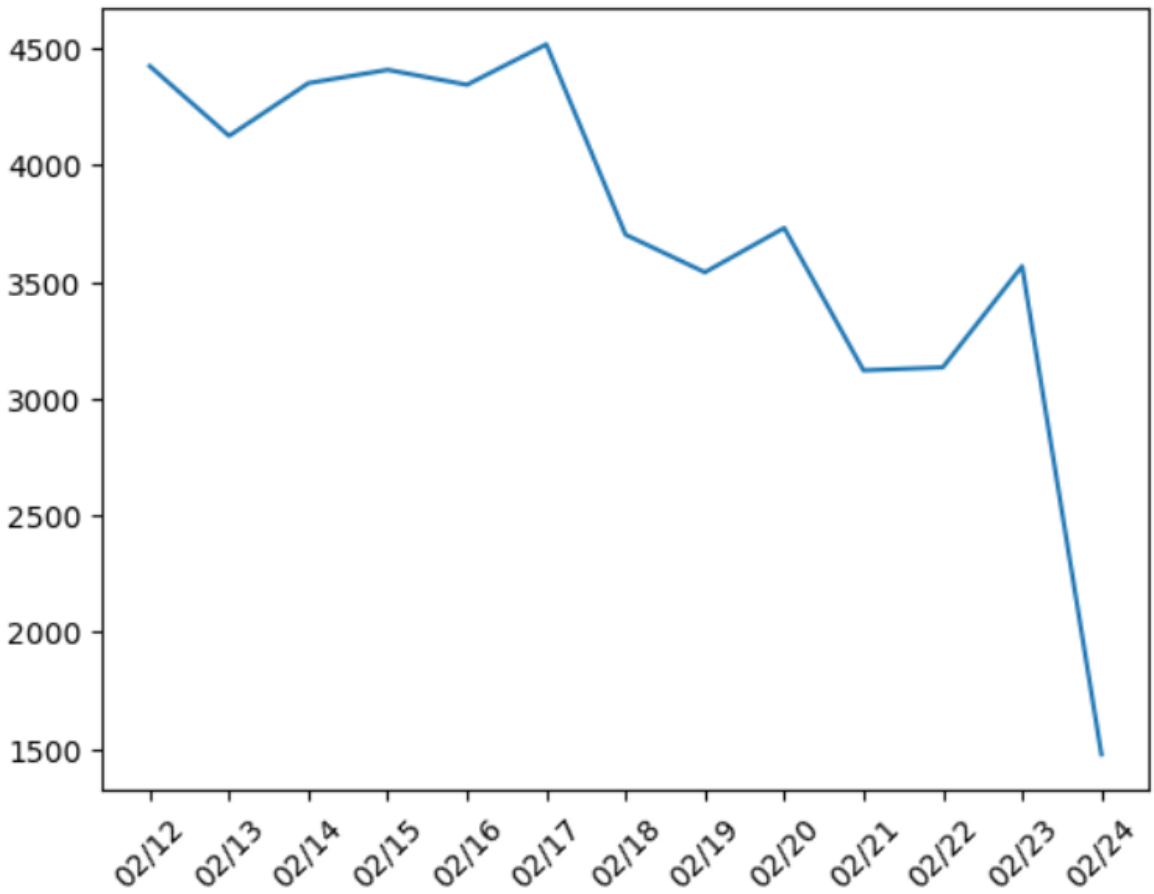
Finally, show a trend line of Tweets for the entire month.

We see that the number of Tweets has a general trend downward as time goes on.

```
import matplotlib.pyplot as plt

value_counts = raw_tweets['date_bin'].value_counts().sort_index()
plt.plot(value_counts)
plt.xticks(rotation=45)
plt.show()
```

The X axis below shows the days in February, while the Y axis shows the number of Tweets. From this line graph, we can see the trend of Tweets throughout February.



Question 2

Find all the hashtags in the Twitter data and do some analysis.

1. What were the top 5 hashtags used in this data?
2. How many individual unique hashtags are used?
3. What is the min, max, and average number of hashtags used per Tweet?

The next set of questions have to do with hashtags used in the Tweets. First, some processing to extract the hashtags from the content of the Tweets:

```
# first, take the 'content' column and make it all lowercase so that #Farmersprotest
# and #farmersprotest will be considered the same hashtag.
raw_tweets['lowercase'] = raw_tweets['content'].str.lower()

# Extract hashtags from 'content' column
raw_tweets['hashtags'] = raw_tweets['lowercase'].str.findall(r'#\w+')
```

```
raw_tweets['hashtags'].value_counts()
```

```
[#farmersprotest]                                                                                  12454
[#pagdi_sambhal_jatta, #farmersprotest]                                                              706
[#farmersmakeindia, #farmersprotest]                                                                 674
[#msplawforallcrops, #farmersprotest]                                                                613
[#farmersprotest, #mahapanchayatrevolution]                                                          608
                                                                                                    ...
[#india, #hindutvafacism, #warcrimes, #kashmir, #farmersprotest, #women, #girls, #freekashmir, #metoo]    1
[#farmersprotest, #nepotism]                                                                           1
[#disharavi, #nikitajacob, #shantanumuluk, #farmersprotest, #farmersmakelndia]                         1
[#farmersprotest, #lpg_petrol_loot]                                                                    1
[#spinelessscelebs, #antinationalbollywood, #farmersprotest, #bjpagainstfarmers]                       1
Name: hashtags, Length: 13230, dtype: int64
```

The code above first converts the content of the Tweets to lowercase to make it easier to search for hashtags and ignore duplicates due to capitalization. For example, we want to count #Farmerprotest and #farmerprotest as the same hashtag, even though the first hashtag has a capital F.

The code then shows the frequency of hashtags used. We will now use this data to answer the questions.

1. What were the top 5 hashtags used in this data?

```
top_5 = raw_tweets['hashtags'].value_counts()[:5]

print('The top 5 hashtags are:')
print(top_5)
```

```
The top 5 hashtags are:
[#farmersprotest]                                  12454
[#pagdi_sambhal_jatta, #farmersprotest]              706
[#farmersmakeindia, #farmersprotest]                 674
[#msplawforallcrops, #farmersprotest]                613
[#farmersprotest, #mahapanchayatrevolution]          608
Name: hashtags, dtype: int64
```

The code above shows the top 5 hashtags used in the Tweets, with #farmersprotest by itself as the most common, with 12,454 occurrences. For context, there are 48,429 Tweets in the entire dataset.

There are other instances where other hashtags are used in conjunction with #farmersprotest, so we will next consider individual hashtags that occur uniquely.

2. How many individual unique hashtags are used?

```
unique_hashtags = []
# extract indivdiual hashtags used
for obj in list(raw_tweets['hashtags']):
    for hashtag in obj:
        unique_hashtags.append(hashtag)
unique_hashtags = set(unique_hashtags) # get only unique hashtags
unique_hashtags = list(unique_hashtags) # cast unique hashtags back into list
unique_hashtags[:10]
```

```
['#shorts',
 '#timesupmodi',
 '#famersmakeindia',
 '#mewsuppasit',
 '#bargarh',
 '#yenicukur',
 '#promote',
 '#2',
 '#communist',
 '#dbz_diary']
```

After pulling out the unique, individual hashtags from the dataset, we can check how many unique hashtags are used across all the Tweets:

```
print('There are {} unique hashtags used.'.format(len(unique_hashtags)))
```

```
There are 7973 unique hashtags used.
```

3. What is the min, max, and average number of hashtags used per Tweet?

```
raw_tweets['hashtags_count'] = raw_tweets['hashtags'].apply(lambda x: len(x))
raw_tweets[['lowercase','hashtags','hashtags_count']]
```

	lowercase	hashtags	hashtags_count
0	the world progresses while the indian police a...	[#modidontsellfarmers, #farmersprotest, #freen...	3
1	#farmersprotest \n#modiignoringfarmersdeaths \...	[#farmersprotest, #modiignoringfarmersdeaths, ...	3
3	@reallyswara @rohini_sgh watch full video here...	[#farmersprotest, #nofarmersnofood]	2
8	@mandeeppunia1 watch full video here https://t...	[#farmersprotest, #nofarmersnofood]	2

The code above creates another column with the number of hashtags used in each Tweet. We can then run some descriptive statistics on this column and answer the question posed above.

```
hashtag_stats = raw_tweets['hashtags_count'].describe()
hashtag_stats
```

```
count    48429.000000
mean         2.816597
std          2.386539
min          1.000000
25%          1.000000
50%          2.000000
75%          3.000000
max         25.000000
Name: hashtags_count, dtype: float64
```

```
ht_min = hashtag_stats[3]
ht_max = hashtag_stats[7]
ht_mean = hashtag_stats[1]

print('The minimum number of hashtags:',ht_min)
print('The maximum number of hashtags:',ht_max)
print('The average number of hashtags:',ht_mean)
```

```
The minimum number of hashtags: 1.0
The maximum number of hashtags: 25.0
The average number of hashtags: 2.816597493237523
```

The output above shows that the minimum number of hashtags used is 1, the maximum is 25, and the average number is 2.82.

Conclusion and key takeaways

Given that the number of Tweets in February using #farmersprotest decreased as the month went on, this suggests that interest in this protest decreased over time, with a spike in interest during the first few days of the protest and a general decline over the next few weeks.

On analysis of hashtags, we see that half of all Tweets contained 2 hashtags (50% quartile), while three-quarters contained 3 hashtags (75% quartile). There were some outliers that contained several more hashtags, the highest number being 25 hashtags in one Tweet.

More could be done to analyze the content of hashtags used, but it is clear that most users Tweeting about the farmer's protest in India only used 2-3 hashtags in their Tweets.