

```
In [2]: 1 import os, glob
        2 import numpy as np
        3 import pandas as pd
        4 import json
        5 import gc
        6 import matplotlib.pyplot as plt
```

```
In [1]: 1 path = 'ECE219_tweet_data\\'
        2
        3 all_files_tweets = dict()
        4
        5 for filename in glob.glob(os.path.join(path, '*.txt')):
        6     hashtag = filename[(len(path) + len('tweets_#')):-4]
        7     # print(hashtag)
        8     each_file_tweets = list()
        9     with open(filename, encoding="utf-8", mode='r') as f:
       10         for line in f:
       11             json_data = json.loads(line)
       12             each_file_tweets.append(json_data)
       13             f.close()
       14         all_files_tweets[hashtag] = each_file_tweets
       15
       16 gc.collect()
```

```
In [3]: 1 all_files_tweets['nfl1'][0].keys()
```

```
Out[3]: dict_keys(['firstpost_date', 'title', 'url', 'tweet', 'author', 'original_author', 'citation_date', 'metrics', 'highlight', 'type', 'citation_url'])
```

Question 9

Question 9.1

```
In [54]: 1 print('Average number of tweets per hour:')
2 for hashtag in all_files_tweets.keys():
3     tweet_time = [i['citation_date'] for i in all_files_tweets[hashtag]]
4     tweet_time = pd.Series(tweet_time)
5     tweet_time = pd.to_datetime(tweet_time, unit = 's')
6     first_tweet_time = tweet_time.min()
7     last_tweet_time = tweet_time.max()
8     time_range = (last_tweet_time - first_tweet_time)
9     time_hours = round(time_range.days*24 + time_range.seconds/3600)
10    average_tweet_per_hour = len(all_files_tweets[hashtag]) / time_hours
11    print(hashtag + ': ', average_tweet_per_hour)
```

Average number of tweets per hour:

gohawks: 292.598615916955
 gopatriots: 40.95993031358885
 nfl: 396.97103918228277
 patriots: 750.6320272572402
 sb49: 1277.7474226804125
 superbowl: 2071.353242320819

```
In [14]: 1 print('Average number of followers per tweet:')
2 for hashtag in all_files_tweets.keys():
3     followers = [i['author']['followers'] for i in all_files_tweets[hashtag]]
4     average_followers_per_tweet = np.mean(followers)
5     print(hashtag + ': ', average_followers_per_tweet)
```

Average number of followers per tweet:

gohawks: 2217.9237355281984
 gopatriots: 1427.2526051635405
 nfl: 4662.37544523693
 patriots: 3280.4635616550277
 sb49: 10374.160292019487
 superbowl: 8814.96799424623

```
In [12]: 1 print('Average number of retweet per tweet:')
2 for hashtag in all_files_tweets.keys():
3     retweets = [i['metrics']['citations']['total'] for i in all_files_tweets[hashtag]]
4     average_retweets_per_tweet = np.mean(retweets)
5     print(hashtag + ': ', average_retweets_per_tweet)
```

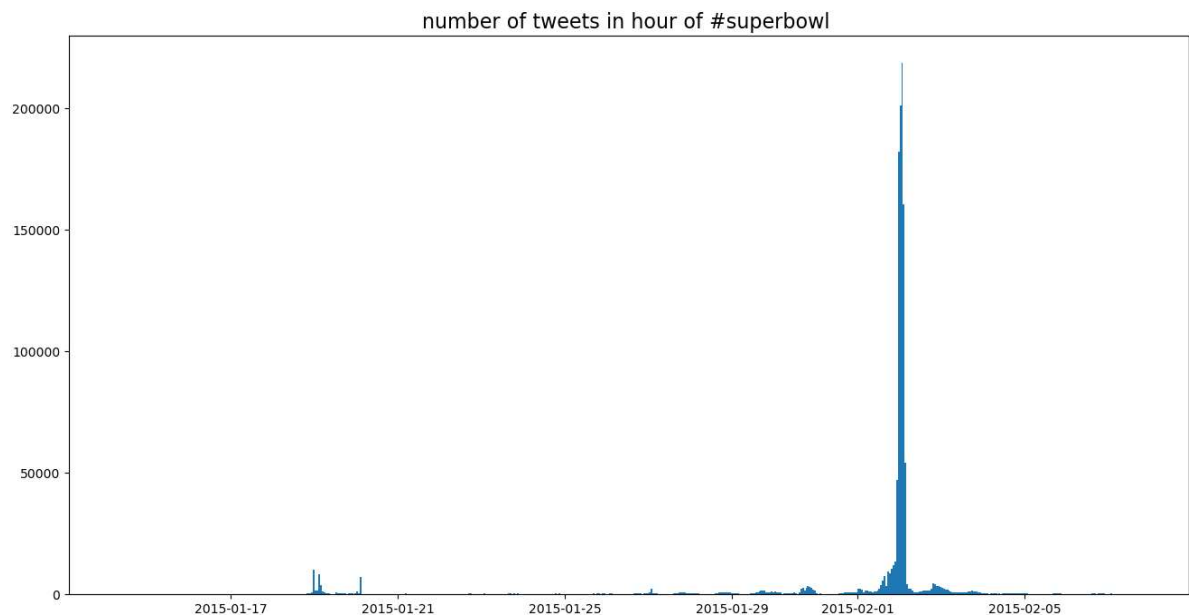
Average number of retweet per tweet:

gohawks: 2.0132093991319877
 gopatriots: 1.4081919101697078
 nfl: 1.5344602655543254
 patriots: 1.7852871288476946
 sb49: 2.52713444111402
 superbowl: 2.3911895819207736

Question 9.2

```
In [116]: 1 superbowl_post_time = [i['citation_date'] for i in all_files_tweets['superbowl']]
2 superbowl_bins = np.arange(min(superbowl_post_time), max(superbowl_post_time), 3600)
3 superbowl_post_time = pd.to_datetime(superbowl_post_time, unit = 's')
4 superbowl_bins = pd.to_datetime(superbowl_bins, unit = 's')
5
6 nfl_post_time = [i['citation_date'] for i in all_files_tweets['nfl']]
7 nfl_bins = np.arange(min(nfl_post_time), max(nfl_post_time) + 1, 3600)
8 nfl_post_time = pd.to_datetime(nfl_post_time, unit = 's')
9 nfl_bins = pd.to_datetime(nfl_bins, unit = 's')
```

```
In [125]: 1 fig, ax = plt.subplots()
2 fig.set_size_inches(16, 8, forward=True)
3 ax.hist(superbowl_post_time, bins= superbowl_bins)
4 plt.title('number of tweets by hour of #superbowl ', fontsize = 16)
5 plt.show()
```



```
In [121]: 1 fig, ax = plt.subplots()
2 fig.set_size_inches(16, 8, forward=True)
3 ax.hist(nfl_post_time, bins= nfl_bins)
4 plt.title('number of tweets by hour of #nfl ', fontsize = 16)
5 plt.show()
```

