initial_interaction

November 13, 2020

1 EDA

This notebook will be going through an overview of the tweets that our team has collected. These tweets were downloaded from a Git repository hosted by the PanaceaLab at Georgia Tech University. They have already filtered these tweets to only those pertaining to the novel Coronavirus.

First let us load the tweets:

We have created a Tweet_Dataset class to easily interact with the tweets. For example there are:

```
910798 tweets in the dataset
```

And this is that a tweet looks like:


```
{'created_at': datetime.datetime(2020, 3, 22, 4, 1, 29,
tzinfo=datetime.timezone.utc),
 'id': 1241575703012716544,
 'id_str': '1241575703012716544',
 'full_text': '327 #Covid19 positive cases in India, \n\nMaharashtra
-64\n\nKerala -52\n\nDelhi -26\n\nUttar Pradesh- 26\n\nRajasthan
-23\n\nTelangana -21\n\nTN - 6\n\nhttps://t.co/ItngnPEJjh
https://t.co/6EGM2NK3vp',
 'truncated': False,
 'display_text_range': [0, 160],
 'entities': {'hashtags': [{'text': 'Covid19', 'indices': [4, 12]}],
  'symbols': [],
  'user_mentions': [],
  'urls': [{'url': 'https://t.co/ItngnPEJjh',
    'expanded_url': 'https://covidout.in/',
    'display_url': 'covidout.in',
    'indices': [137, 160]}],
  'media': [{'id': 1241575692866711552,
    'id_str': '1241575692866711552',
    'indices': [161, 184],
    'media_url': 'http://pbs.twimg.com/media/ETr2g7cUcAAaQPM.jpg',
    'media_url_https': 'https://pbs.twimg.com/media/ETr2g7cUcAAaQPM.jpg',
    'url': 'https://t.co/6EGM2NK3vp',
    'display_url': 'pic.twitter.com/6EGM2NK3vp',
    'expanded_url':
'https://twitter.com/cinema_war/status/1241575703012716544/photo/1',
```

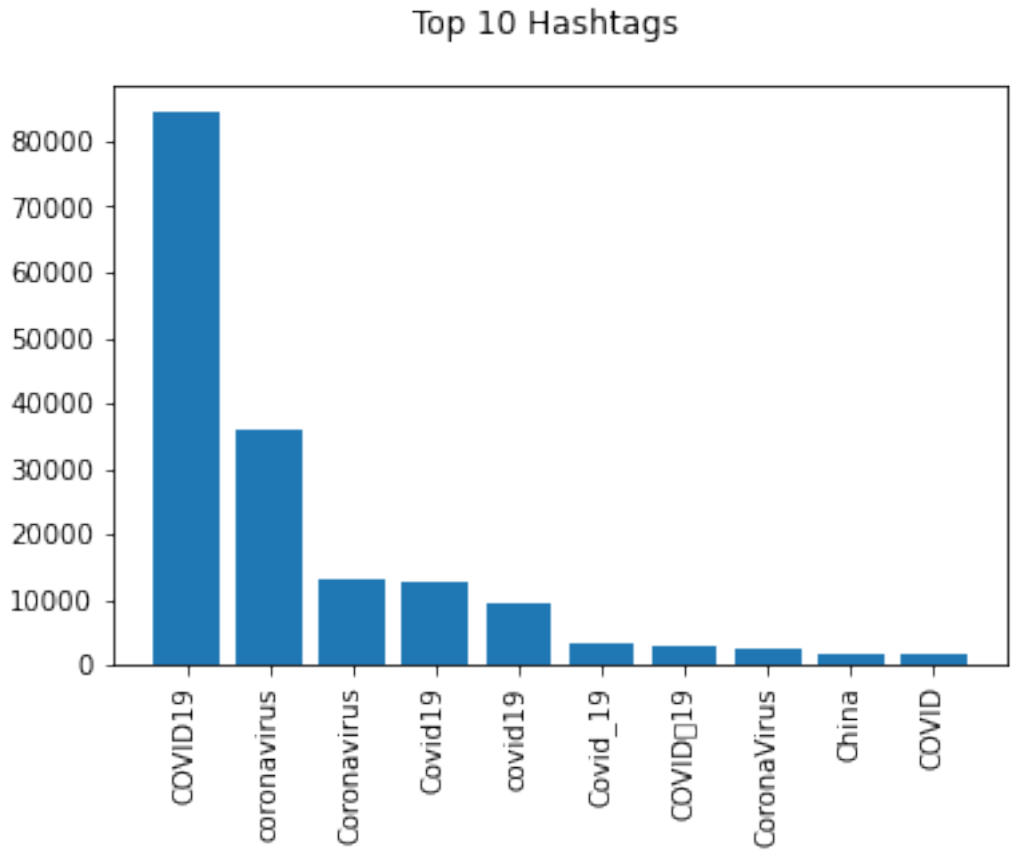
```
    'type': 'photo',
    'sizes': {'large': {'w': 980, 'h': 1060, 'resize': 'fit'},
     'thumb': {'w': 150, 'h': 150, 'resize': 'crop'},
     'small': {'w': 629, 'h': 680, 'resize': 'fit'},
     'medium': {'w': 980, 'h': 1060, 'resize': 'fit'}}}]},
 'extended_entities': {'media': [{'id': 1241575692866711552,
    'id_str': '1241575692866711552',
    'indices': [161, 184],
    'media_url': 'http://pbs.twimg.com/media/ETr2g7cUcAAaQPM.jpg',
    'media_url_https': 'https://pbs.twimg.com/media/ETr2g7cUcAAaQPM.jpg',
    'url': 'https://t.co/6EGM2NK3vp',
    'display_url': 'pic.twitter.com/6EGM2NK3vp',
    'expanded_url':
'https://twitter.com/cinema_war/status/1241575703012716544/photo/1',
    'type': 'photo',
    'sizes': {'large': {'w': 980, 'h': 1060, 'resize': 'fit'},
     'thumb': {'w': 150, 'h': 150, 'resize': 'crop'},
     'small': {'w': 629, 'h': 680, 'resize': 'fit'},
     'medium': {'w': 980, 'h': 1060, 'resize': 'fit'}},
    'ext_alt_text': None}]},
 'source': '<a href="http://twitter.com/download/android" rel="nofollow">Twitter
for Android</a>',
 'in_reply_to_status_id': None,
 'in_reply_to_status_id_str': None,
 'in_reply_to_user_id': None,
 'in_reply_to_user_id_str': None,
 'in_reply_to_screen_name': None,
 'user': {'id': 1001592067,
  'id_str': '1001592067',
  'name': 'Tamil Cinema War',
  'screen_name': 'cinema_war',
  'location': 'Chennai, India',
  'description': 'Cinema News 24/7 | Movie Reviews | Galleries,Trailers | For
Promotions DM Me.\n\nEllarum Nalla Iruppom',
  'url': 'https://t.co/1cHAOKzql0',
  'entities': {'url': {'urls': [{'url': 'https://t.co/1cHAOKzql0',
      'expanded_url': 'http://www.ndmsnow.com',
      'display_url': 'ndmsnow.com',
      'indices': [0, 23]}]},
   'description': {'urls': []}},
  'protected': False,
  'followers_count': 14354,
  'friends_count': 178,
  'listed_count': 40,
  'created_at': 'Mon Dec 10 12:37:36 +0000 2012',
  'favourites_count': 1254,
  'utc_offset': None,
```

```
  'time_zone': None,
  'geo_enabled': False,
  'verified': False,
  'statuses_count': 34124,
  'lang': None,
  'contributors_enabled': False,
  'is_translator': False,
  'is_translation_enabled': False,
  'profile_background_color': 'C0DEED',
  'profile_background_image_url':
'http://abs.twimg.com/images/themes/theme1/bg.png',
  'profile_background_image_url_https':
'https://abs.twimg.com/images/themes/theme1/bg.png',
  'profile_background_tile': True,
  'profile_image_url':
'http://pbs.twimg.com/profile_images/1260890794203938819/u5QEcpYm_normal.jpg',
  'profile_image_url_https':
'https://pbs.twimg.com/profile_images/1260890794203938819/u5QEcpYm_normal.jpg',
  'profile_banner_url':
'https://pbs.twimg.com/profile_banners/1001592067/1596094839',
  'profile_image_extensions_alt_text': None,
  'profile_banner_extensions_alt_text': None,
  'profile_link_color': '0084B4',
  'profile_sidebar_border_color': 'FFFFFF',
  'profile_sidebar_fill_color': 'DDEEF6',
  'profile_text_color': '333333',
  'profile_use_background_image': True,
  'has_extended_profile': True,
  'default_profile': False,
  'default_profile_image': False,
  'following': False,
  'follow_request_sent': False,
  'notifications': False,
  'translator_type': 'none'},
 'geo': None,
 'coordinates': None,
 'place': None,
 'contributors': None,
 'is_quote_status': False,
 'retweet_count': 0,
 'favorite_count': 0,
 'favorited': False,
 'retweeted': False,
 'possibly_sensitive': False,
 'lang': 'in'}
```


As you see, the JSON format of the tweet allows us to easily access the metadata for each tweet.

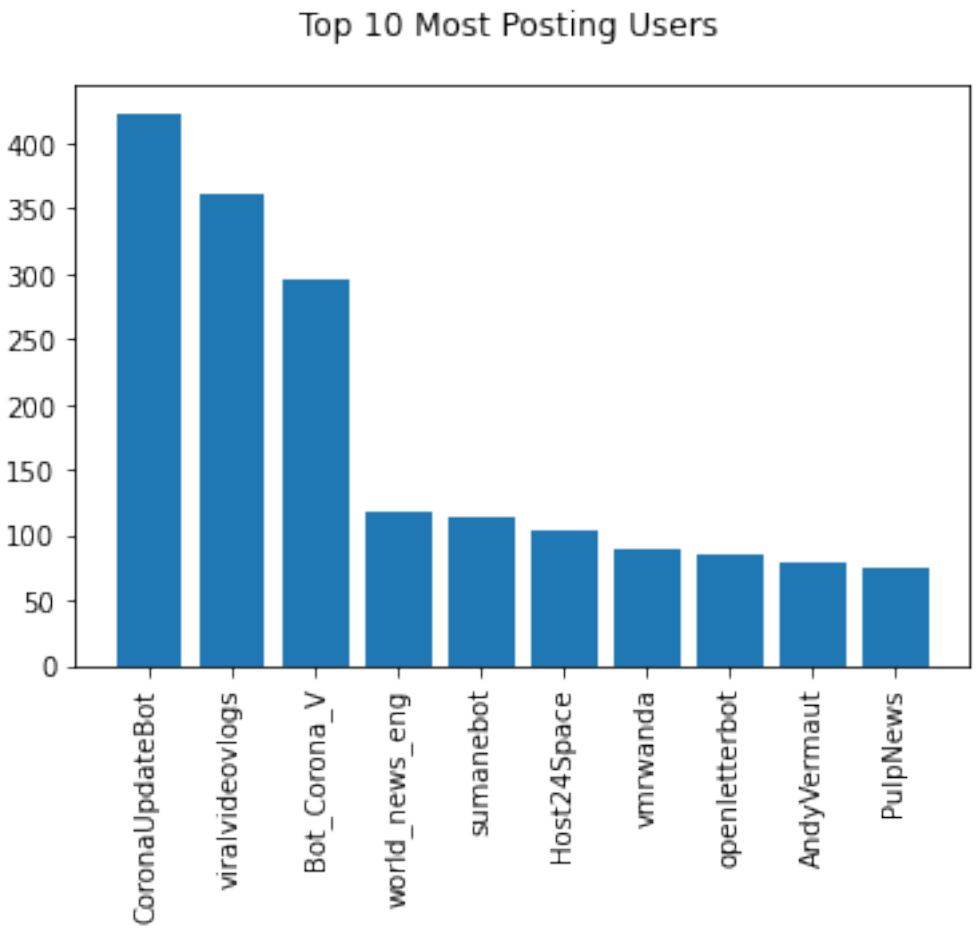
1.1 Top 10 Hashtags

One of the most important pieces of information contained in the tweets are the hashtags. Hashtags are a special kind of data because they serve as a sort of classification by themselves. Here, we look at the top 10 hashtags that were present in the dataset.



1.2 Most Posting Users

Now we shall do the same but for the number of times each user posts.



Immediately we notice something interesting. The users who post the most frequently are bots. They seem to make up a large proportion of the Top 10 as well, with 40% of the top ten users containing the phrase 'bot' in their names. Of course, some of the other users may be bots as well, but just without the word 'bot' in their screen name. This definitely warrants further investigation.

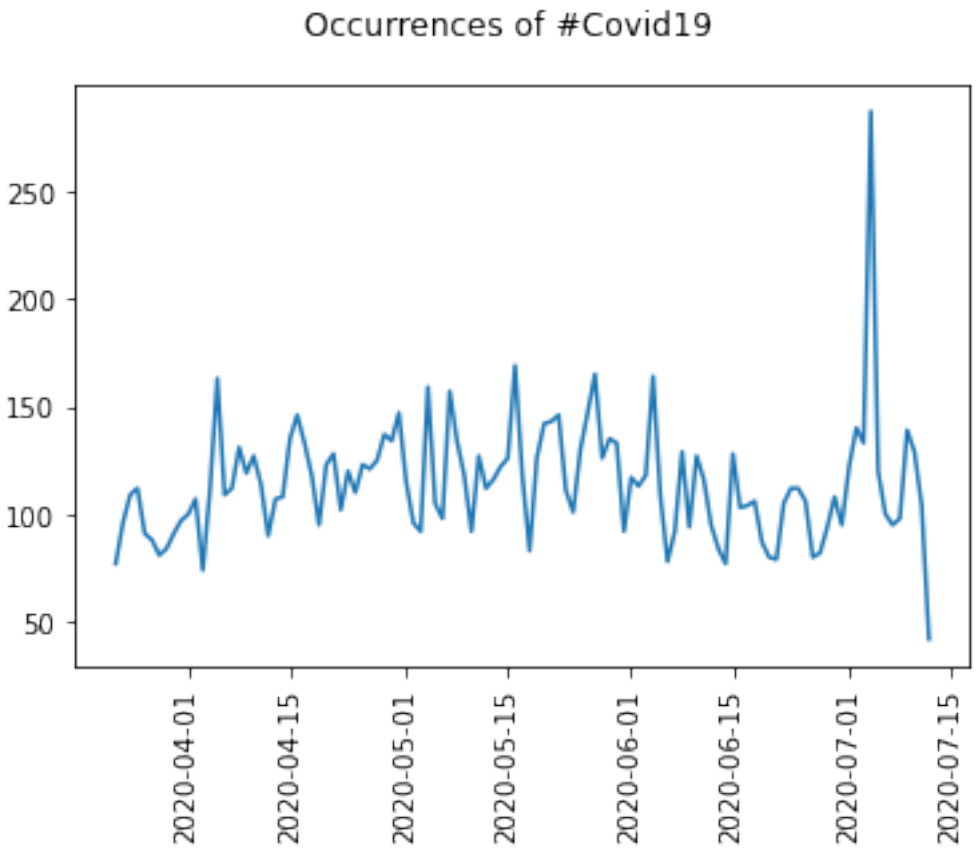
1.3 Hashtag Counts by Day

In this section we will select, by hand, 3 hashtags for the science set and 3 for the conspiracy set. Note that this was based off of our biased assumptions and could very well not be the case. We will investigate how the usage of these hashtags changed throughout the dataset's timespan.

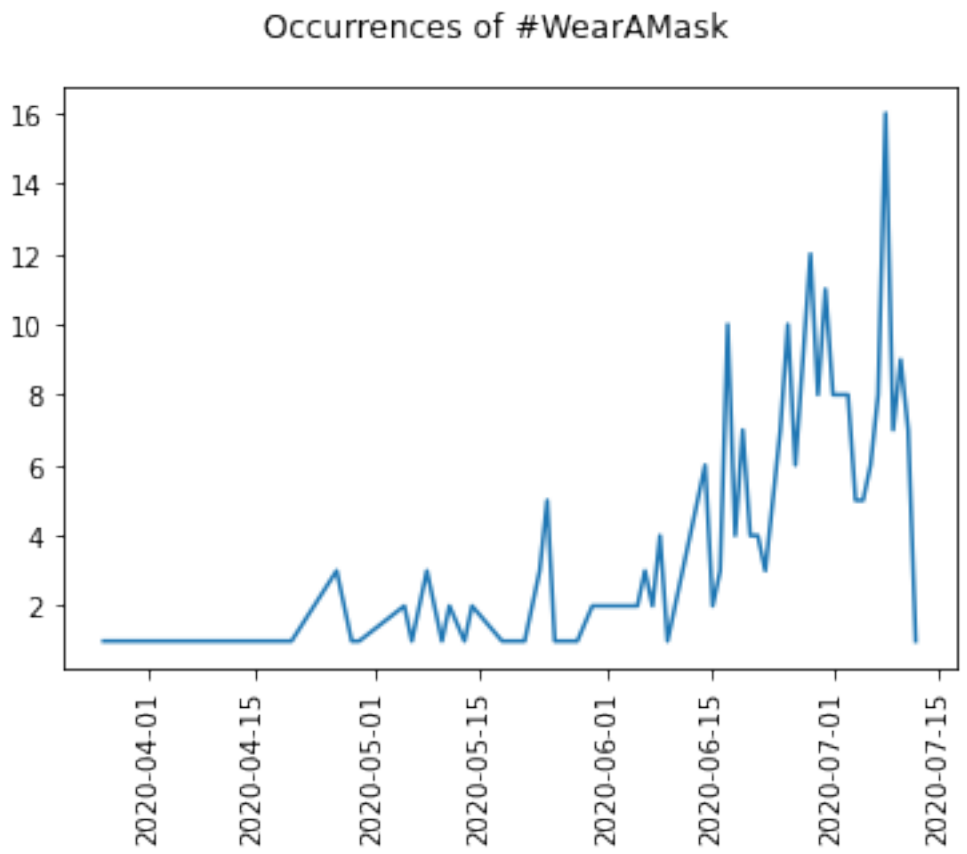
1.3.1 Science Hashtag Usages

We chose to associate #Covid19, #WearAMask, and #SocialDistancing with being categorized as science based tweets. Below we portray the usage of the these hashtags throughout the days contained in the dataset we collected.

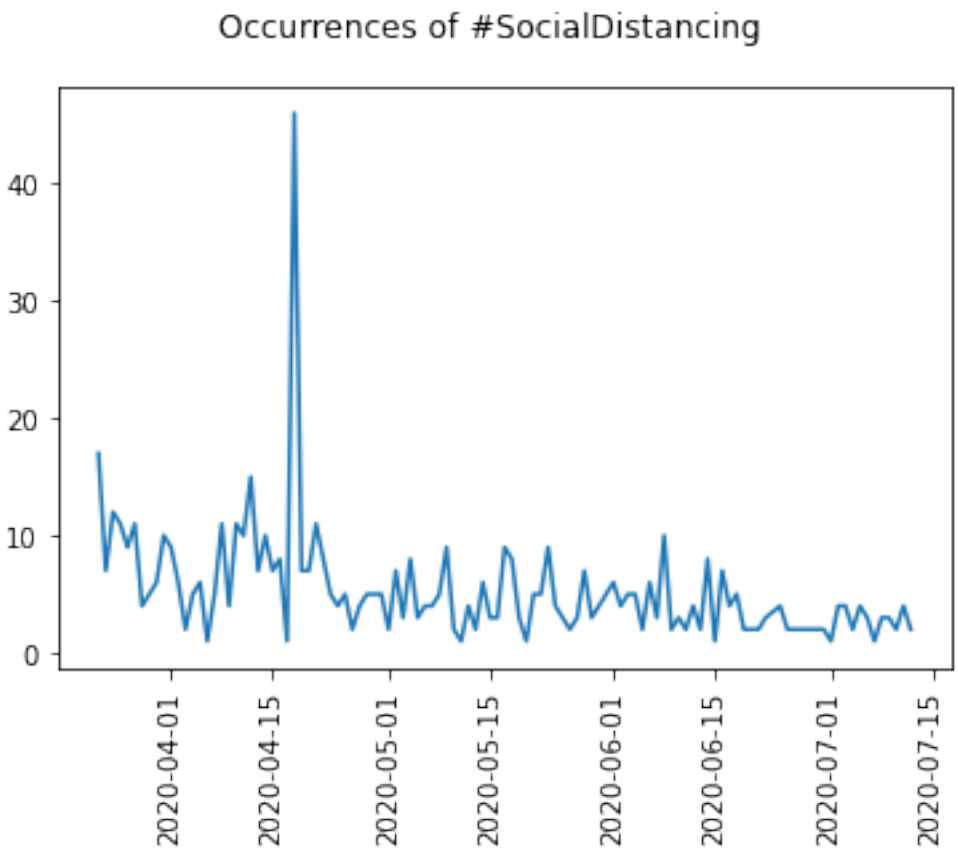
```
[<matplotlib.lines.Line2D at 0x7f45185c2050>]
```




```
[<matplotlib.lines.Line2D at 0x7f43b38ad910>]
```



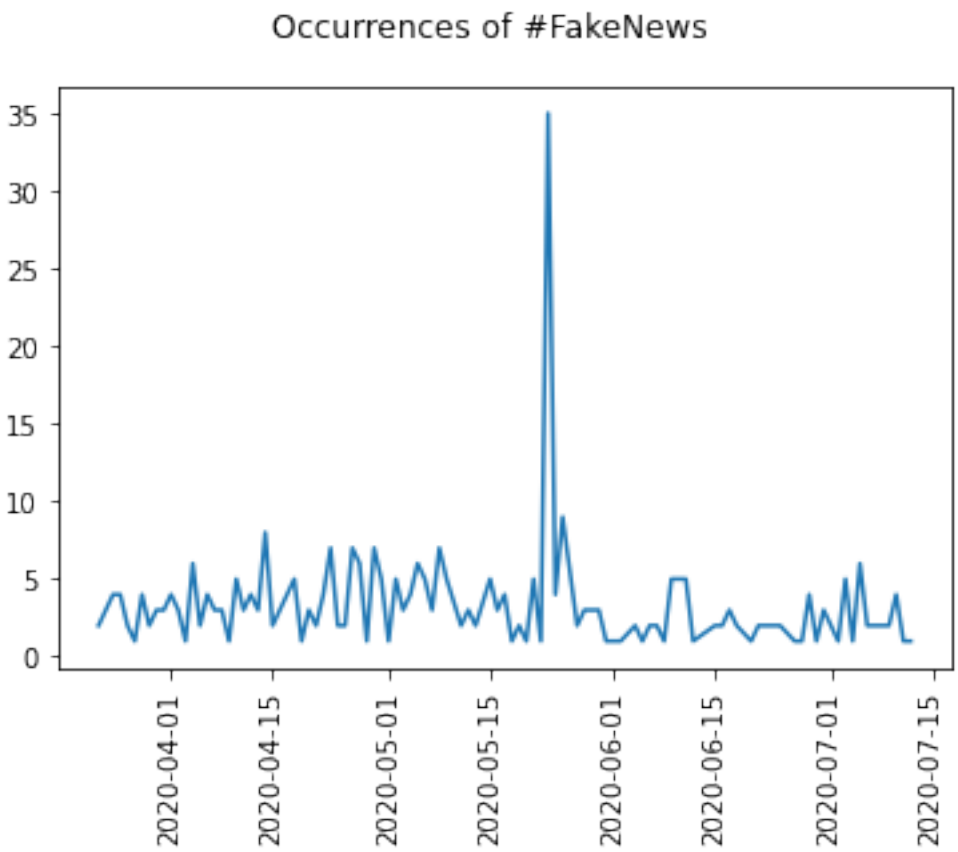

```
[<matplotlib.lines.Line2D at 0x7f43b38a1190>]
```



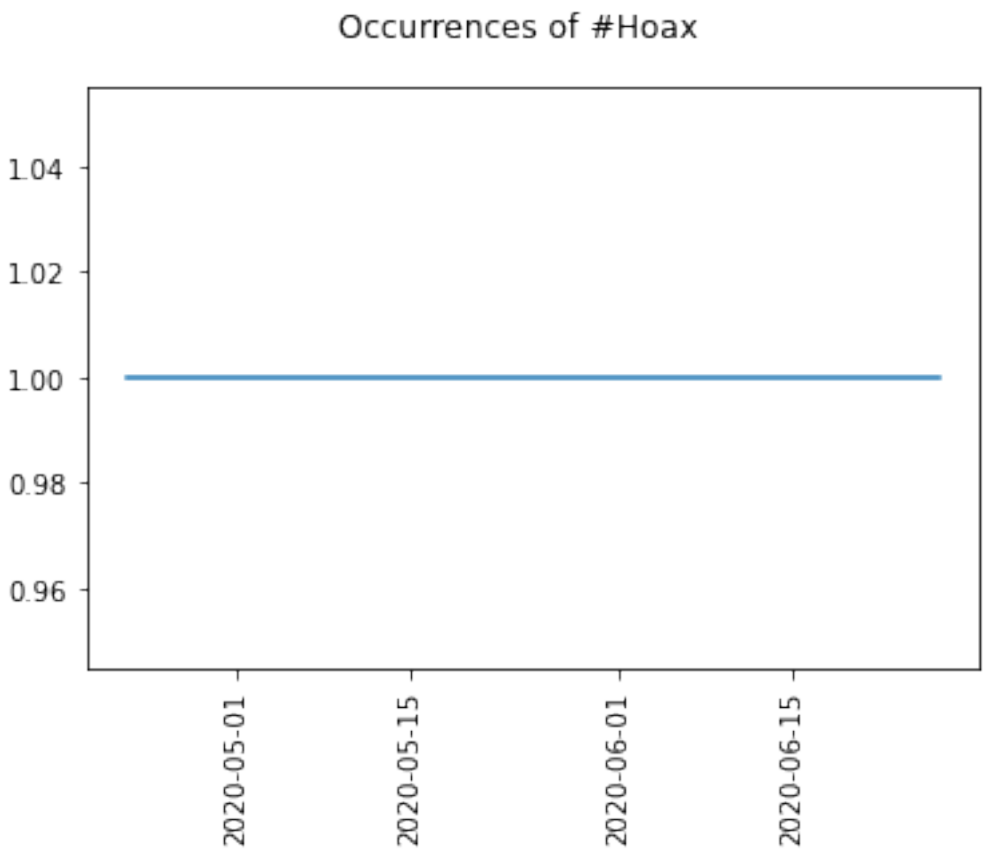
1.3.2 Misinformation Hashtag Usages

Categorizing hashtags as misinformation based was subjective, just as it was for the former category. We chose #FakeNews, #Hoax, and #ChinaVirus to be the hashtags associated with tweets that were not based on fact. Their daily usage plots are below.

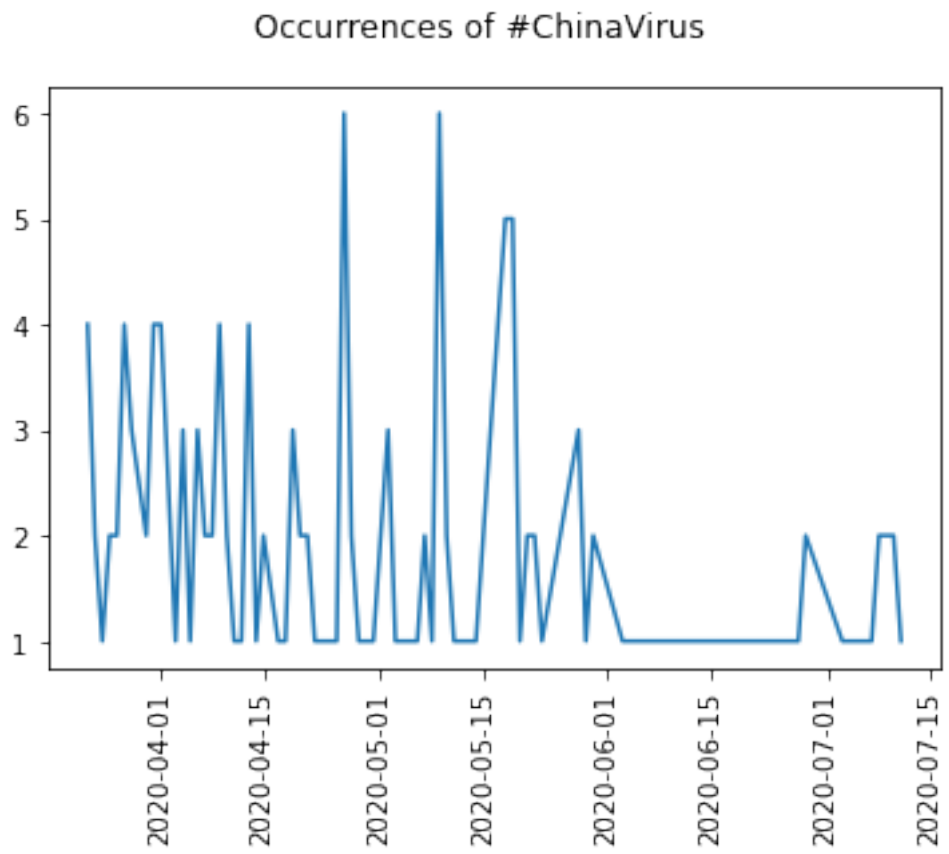
```
[<matplotlib.lines.Line2D at 0x7f43b5fccc90>]
```




```
[<matplotlib.lines.Line2D at 0x7f45185c2490>]
```




```
[<matplotlib.lines.Line2D at 0x7f43b3839290>]
```


Occurrences of #ChinaVirus
6
5
4
3
2
1
2020-04-01
2020-04-15
2020-05-01
2020-05-15
2020-06-01
2020-06-15
2020-07-01
2020-07-15