# Tutorial: Introduction to Data Science with Python
# #brexit Analysis

**Data Source**: Twitter, using stream API
**Data type**: Tweets regarding brexit
**Number of tweets collected**: 5000

**Libraries Used**:

- PyMongo: Mongo DB client for Python
  - https://api.mongodb.com/python/current/
- Tweepy: Python library to access Twitter API
  - https://tweepy.readthedocs.io/en/v3.5.0/

**Pre-processing:**
Just tracked twitter with #brexit to figure out what other hashtags might be relevant. The final hashtags to be tracked were:

- 'brexit'
- 'theresamay'
- 'stopbrexit',
- 'brexitdeal',
- 'brexitshambles'

To setup MongoDB collection with the fetched tweets from the API, use the following:

```
$ mongoimport -d analysis -c brexit  --file brexit.json
```

**Analysis**

1. **Summary: Top 20 hashtags**
2. **Summary: Mentions of Theresa May**
3. **Summary: Counts of the 'favorites' on tweets**
4. **Comparison: Types of tweets - original vs retweeted**
5. **Comparison: Types of tweet content - text vs audio vs video**

**Description:**
Stream API is used to fetch the tweets for the specified hashtags, and the data is stored in a MongoDB collection. The script 'fetchTweet.py' takes a secret file with the API tokens, and the tweepy API requests to fetch the data. Pymongo is used as a connector to store the data into a collection, so that we can run queries on the data.

The analysis is done using the script 'analysis.py' and it has 5 methods to analyse the five different questions. Part of those are done using Python, and a part of them are done using raw Mongo queries - as it seemed to be faster.

In the following page, we'll see the results of our analysis.

# Top 20 hashtags


*(word cloud generated using: [https://www.wordclouds.com](https://www.wordclouds.com))*

Hashtags with their frequencies are:

| | |
|---|---|
| 573 | brexit |
| 36 | eu |
| 32 | revokearticle50 |
| 27 | indicativevotes |
| 26 | brexitshambles |
| 24 | wtobrexitnow |
| 23 | peoplesvote |
| 21 | brexitstorm |
| 15 | nodeal |
| 14 | nhs |
| 13 | indicativevotes2 |
| 12 | theresamay |
| 12 | singlemarket |
| 10 | stopbrexit |
| 9 | r4today |

| | |
|---|---|
| 9 | twatinahat |
| 8 | brexitbetrayal |
| 8 | uk |
| 7 | brexitchaos |
| 7 | fbpc |

# Mentions of Theresa May

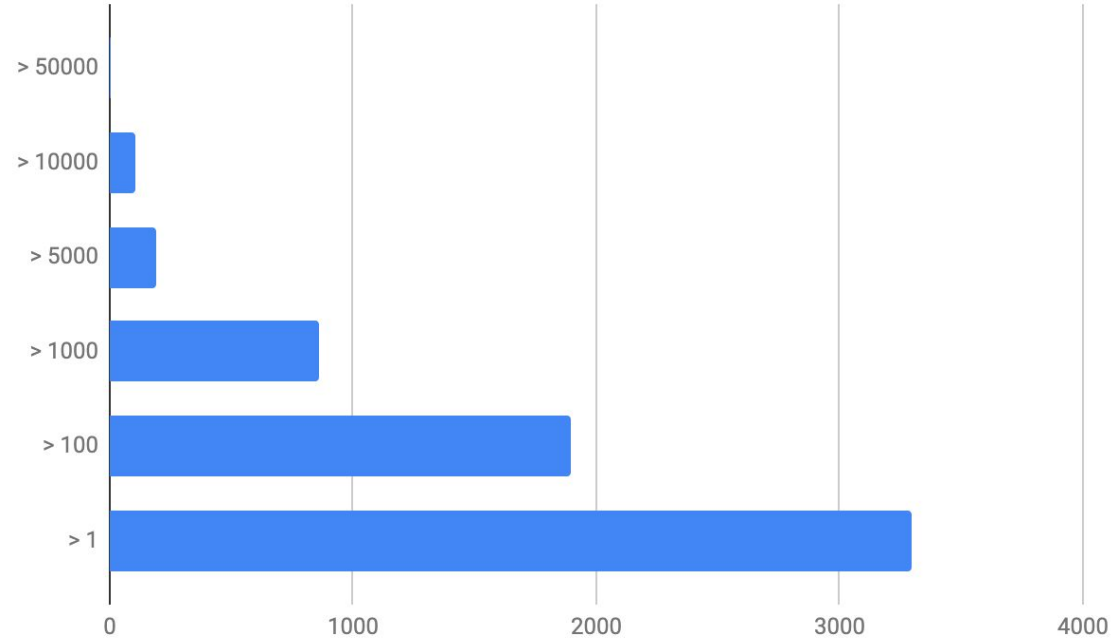
We used a regex pattern to match all the names from “'theresa|theresamay|maytheresa'
**Total number of mentions in 5000 tweets: 153**

# Counts of favorites on tweets

Observartions: It seems like only 3/5000 tweets were favorited more than 50,000 times. Also, 3295 tweets were favorited atleast 1 time

The whole list is here for some numbers:

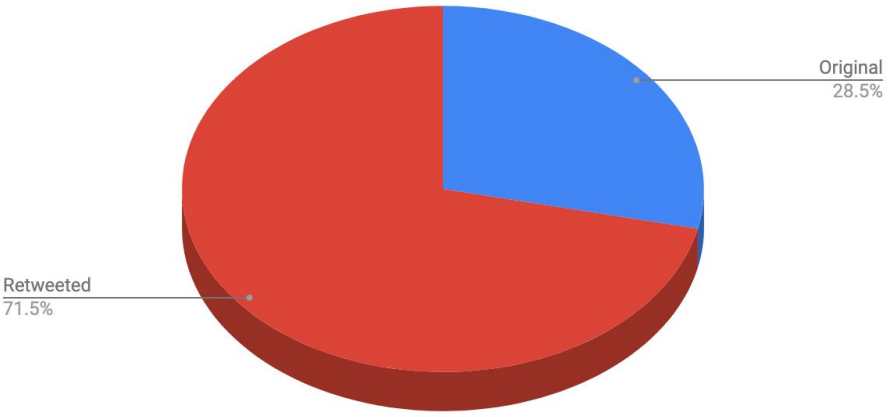
| | |
|---|---|
| > 50000 | 3 |
| > 10000 | 108 |
| > 5000 | 187 |
| > 1000 | 861 |
| > 100 | 1896 |
| > 1 | 3295 |



# Original Tweets vs Retweeted Tweets

**Original: 1424/5000**
**Retweeted: 3576/5000**

On Twitter, a majority of the users retweet what they find interesting, and the behaviour could also beconfirmed from the analysis




# Tweet Types

| | |
|---|---|
| Text | 4,822 |
| Photo | 148 |
| Video | 19 |