Home Work 4

- Aman Joshi (2018201097)

- Link of data : https://drive.google.com/open?id=1BsPhhKkVqjpaW7pukyk3wdARK1QEXqta

```
import requests
import bs4
from bs4 import BeautifulSoup
from requests_oauthlib import OAuth1
import tweepy
from pprint import pprint
import numpy as np
from functools import reduce
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
from collections import Counter
import matplotlib.pyplot as plt
import pickle
import pandas as pd
```

Question 1

```
def get_all_tweets(screen_name):
    all_tweets = []
    auth_params = {
        'app_key': 'app_key',
        'app_secret': 'app_secret',
        'oauth_token': 'oauth_token',
        'oauth_token_secret': 'oauth_token_secret'
    }

    auth = tweepy.OAuthHandler(auth_params['app_key'],
                               auth_params['app_secret'])
    auth.set_access_token(auth_params['oauth_token'],
                          auth_params['oauth_token_secret'])

    api = tweepy.API(auth)
    new_tweets = api.user_timeline(screen_name=screen_name, count=200)
    all_tweets.extend(new_tweets)
    oldest = all_tweets[-1].id - 1
    while len(new_tweets) > 0:
        new_tweets = api.user_timeline(
            screen_name=screen_name, count=200, max_id=oldest)
        all_tweets.extend(new_tweets)
        oldest = all_tweets[-1].id - 1

    return all_tweets
```

```
bjp_tweets = get_all_tweets('BJP4India')
```

```
with open('BJP_tweets.pickle', 'wb') as output_file:
    pickle.dump(bjp_tweets, output_file)
with open('BJP_tweets.pickle', 'rb') as input_file:
    bjp_tweets = pickle.load(input_file)
```

1-a Tweets with most likes

Tweet with id 1110886268408233984 has most number of likes

```
tweet_with_most_like_id = bjp_tweets[np.argmax(
    np.array([tweet.favorite_count for tweet in bjp_tweets]))].id
```

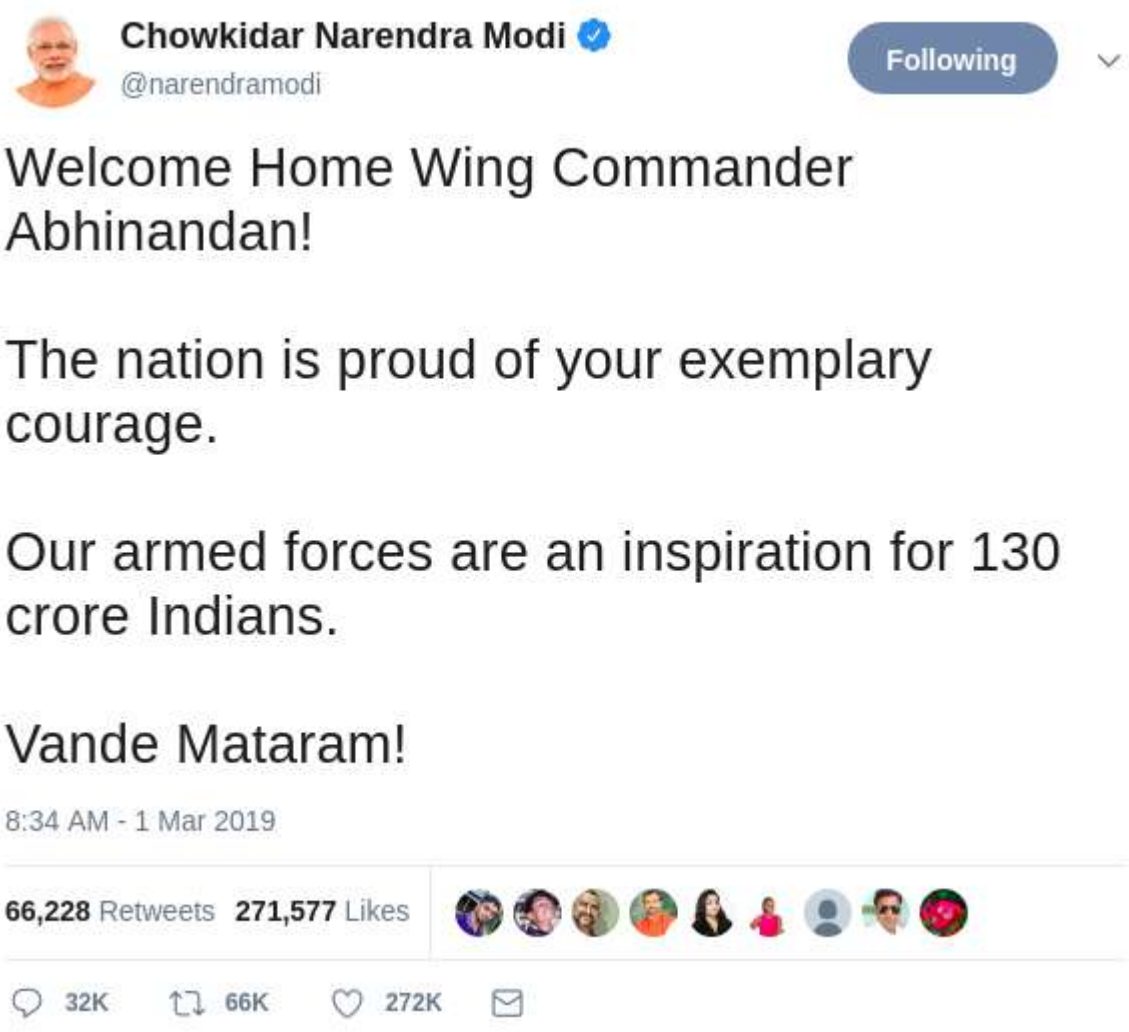
```
tweet_with_most_like_id
```

```
1110886268408233984
```

Q1-b Tweet with most number of retweets.

Tweet with id 1101524057353342981 has most number of retweets

```
tweet_with_most_retweets = bjp_tweets[np.argmax(
    np.array([tweet.retweet_count for tweet in bjp_tweets]))].id
```



Q1-c Top 5 most popular tweets

I calculated popularity of tweet as sum of favourite count and two times retweet count. Id of msot popular tweets are as follow:

- 1101524057353342981
- 1106767552351559680
- 1114495595421376512
- 1110886268408233984
- 1110802295527079936

```
def popularity(tweet):
    return tweet.favorite_count + tweet.retweet_count * 2
```

```
most_popular_tweets = sorted(
    bjp_tweets, key=lambda x: popularity(x), reverse=True)
```

```
popular_ids = [tweet.id for tweet in most_popular_tweets[:5]]
print(popular_ids)
```

```
[1101524057353342981, 1106767552351559680, 1114495595421376512,
1110886268408233984, 1110802295527079936]
```

Q1-d Top 5 msot used hashtags by BJP's official handle:

Top 5 most used hashtags by BJP are as follow:

- BJPSankalpPatr2019
- MainBhiChowkidar
- BJPVijaySankalpBikeRally
- DeshKeLiyeModi
- IsBaarNaMoPhirSe

```
hashtags = reduce(lambda a, b: a + b,
                  [tweet.entities['hashtags'] for tweet in bjp_tweets])
hashtags, count = np.unique([i['text'] for i in hashtags],
return_counts=True)
hashtags = sorted(zip(hashtags, count), key=lambda x: x[1], reverse=True)
most_popular_hashtags = [hashtag[0] for hashtag in hashtags[:5]]
```

```
pprint(most_popular_hashtags)
```

```
['BJPSankalpPatr2019',
 'MainBhiChowkidar',
 'BJPVijaySankalpBikeRally',
 'DeshKeLiyeModi',
 'IsBaarNaMoPhirSe']
```

Question 2

Collect 3200 most recent tweets by Indian National Congress's official twitter handle

```
inc_tweets = get_all_tweets('INCIndia')
```

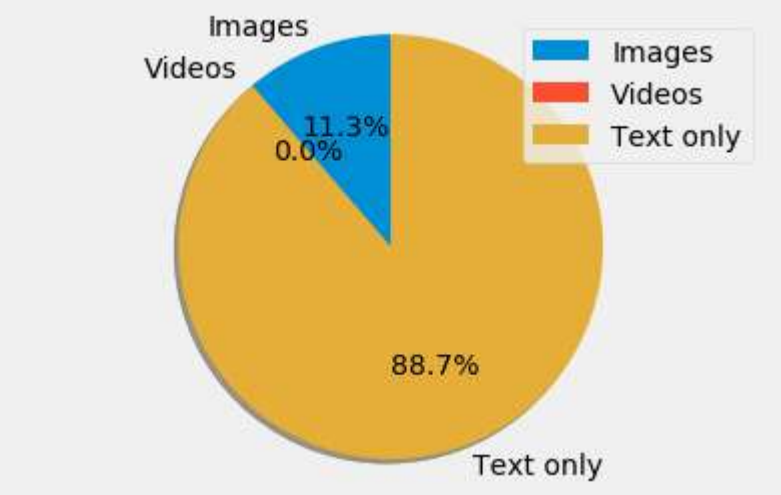
```
with open('INC_tweets.pickle', 'wb') as output_file:
    pickle.dump(inc_tweets, output_file)
```

```
with open('INC_tweets.pickle', 'rb') as input_file:
    inc_tweets = pickle.load(input_file)
```

Q2-a How many tweets consist of images

```
image_count = text_only_count = video_count = 0
for status in inc_tweets:
    if 'media' in status.entities:
        if (status.entities['media'][0]['type'] == "photo"):
            image_count += 1
        elif (status.entities['media'][0]['type'] == "video"):
            video_count += 1
    else:
        text_only_count += 1

counts = [image_count, video_count, text_only_count]
value = ['Images', 'Videos', 'Text only']
fig1, ax1 = plt.subplots()
ax1.pie(counts, labels=value, autopct='%1.1f%%', shadow=True,
startangle=90)
ax1.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a
circle.
plt.legend(loc="best")
plt.style.use('fivethirtyeight')
plt.show()
```



Q2-b Plot word cloud of 1000 most popular tweets

```
most_popular_tweets = sorted(
    inc_tweets, key=lambda x: popularity(x), reverse=True)[:1000]
corpus = reduce(lambda a, b: a + b,
                [tweet.text for tweet in most_popular_tweets]).lower()
for c in [
        '.', ',', '"', "'", '?', '!', '/', '\\', ':', ';', '(', ')', '[',
']',
```

```
        '{', '}', '&', '@'
]:
    corpus = corpus.replace(c, "")
corpus = corpus.split()
```

```
stopwords = set(STOPWORDS)
corpus = [i for i in corpus if i not in stopwords]
wordcloud = WordCloud(background_color="white", stopwords=stopwords)
wordcloud_dict = Counter(corpus)
wordcloud.generate_from_frequencies(wordcloud_dict)
plt.figure(figsize=(15, 8))
plt.imshow(wordcloud)
plt.axis("off")
plt.show()
```

Q2-c List top 5 most used hashtags by INC's official handle

- RafaleSca
- JanKiBaat
- OneNationBillionIdeas
- ChowkidarChorHai
- LoveNotHate

```
hashtags = reduce(lambda a, b: a + b,
                  [tweet.entities['hashtags'] for tweet in inc_tweets])
hashtags, count = np.unique([i['text'] for i in hashtags],
return_counts=True)
hashtags = sorted(zip(hashtags, count), key=lambda x: x[1], reverse=True)
```

```
most_popular_hashtags = [hashtag[0] for hashtag in hashtags[:5]]
pprint(most_popular_hashtags)
```

```
['RafaleScam',
 'JanKiBaat',
 'OneNationBillionIdeas',
 'ChowkidarChorHai',
 'LoveNotHate']
```

Part 2

Scraping Instagram. I've used Instaloader.

Question 1:

Collect 500 most recent post by Indian National Congress's official Instagram Handle.

```
from instaloader import *
```

```
L = Instaloader()
profile = Profile.from_username(L.context, 'incindia')
```

```
posts = []
for post in profile.get_posts():
    posts.append(post)
    if len(posts) == 500:
        break
```

```
with open('INC_post.pickle', 'wb') as output_file:
    pickle.dump(posts, output_file)
with open('INC_post.pickle', 'rb') as input_file:
    posts = pickle.load(input_file)
```

Q1-a Most liked post

Most liked post has shortcode Bs-IxK3FTIx

```
most_liked_post = posts[np.argmax([post.likes for post in
posts])].shortcode
```

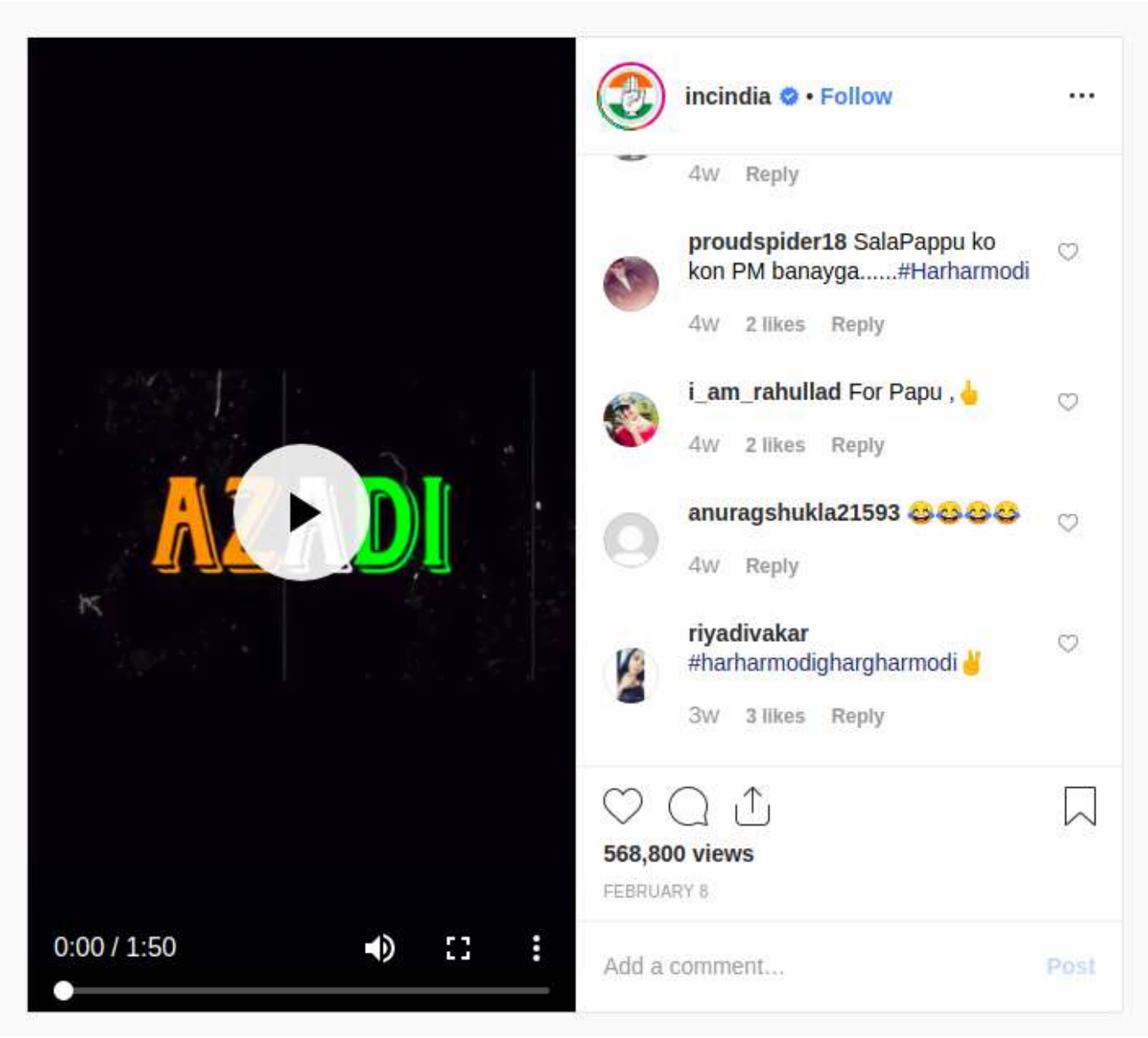
```
print(most_liked_post)
```

Q1-b most commented post

Most commented post has short code BtoT5MoFAL3

```
most_commented_post = posts[np.argmax(
    [post.comments for post in posts[:]])].shortcode
print(most_commented_post)
```

```
BtoT5MoFAL3
```



Q1-c 5 Most popular post

Popularity of post is defined as sum of number of likes and twice the number of comments. Following are most popular posts:

- Bs-IxK3FTIx
- BrmutwOF63Y
- BrQDi4VFI2-
- Bvgd6WZlmsz
- Bt6QHouleGN

```
def popularity_of_post(post):
    return post.likes + post.comments * 2
```

```
most_popular_post = sorted(
    posts, key=lambda x: popularity_of_post(x), reverse=True)
```

```
popular_post_shortcode = [post.shortcode for post in most_popular_post[:5]]
print(popular_post_shortcode)
```

```
['Bs-IxK3FTIx', 'BrmutwOF63Y', 'BrQDi4VFI2-', 'Bvgd6WZlmsz', 'Bt6QHouleGN']
```

Question 2:

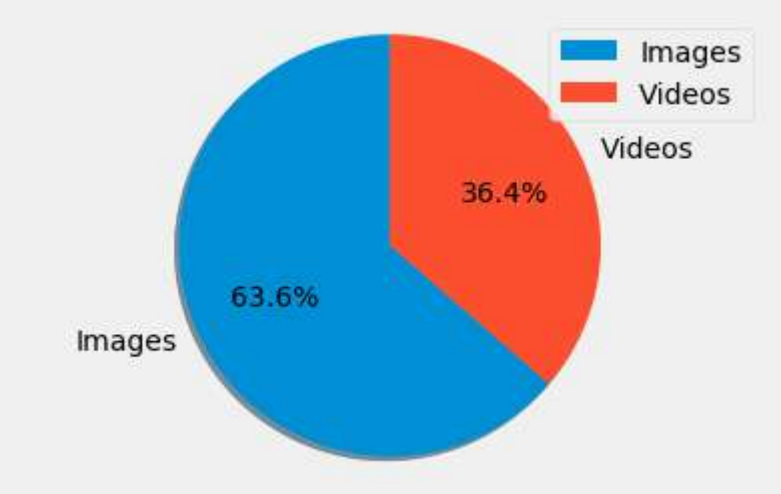
Collect 1,000 most recent posts from the BJP's official Instagram handle.

```
profile = Profile.from_username(L.context, 'BJP4India')
posts = []
for post in profile.get_posts():
    posts.append(post)
    if len(posts) == 1000:
        break
```

```
with open('BJP_post.pickle', 'wb') as output_file:
    pickle.dump(posts, output_file)
with open('BJP_post.pickle', 'rb') as input_file:
    posts = pickle.load(input_file)
```

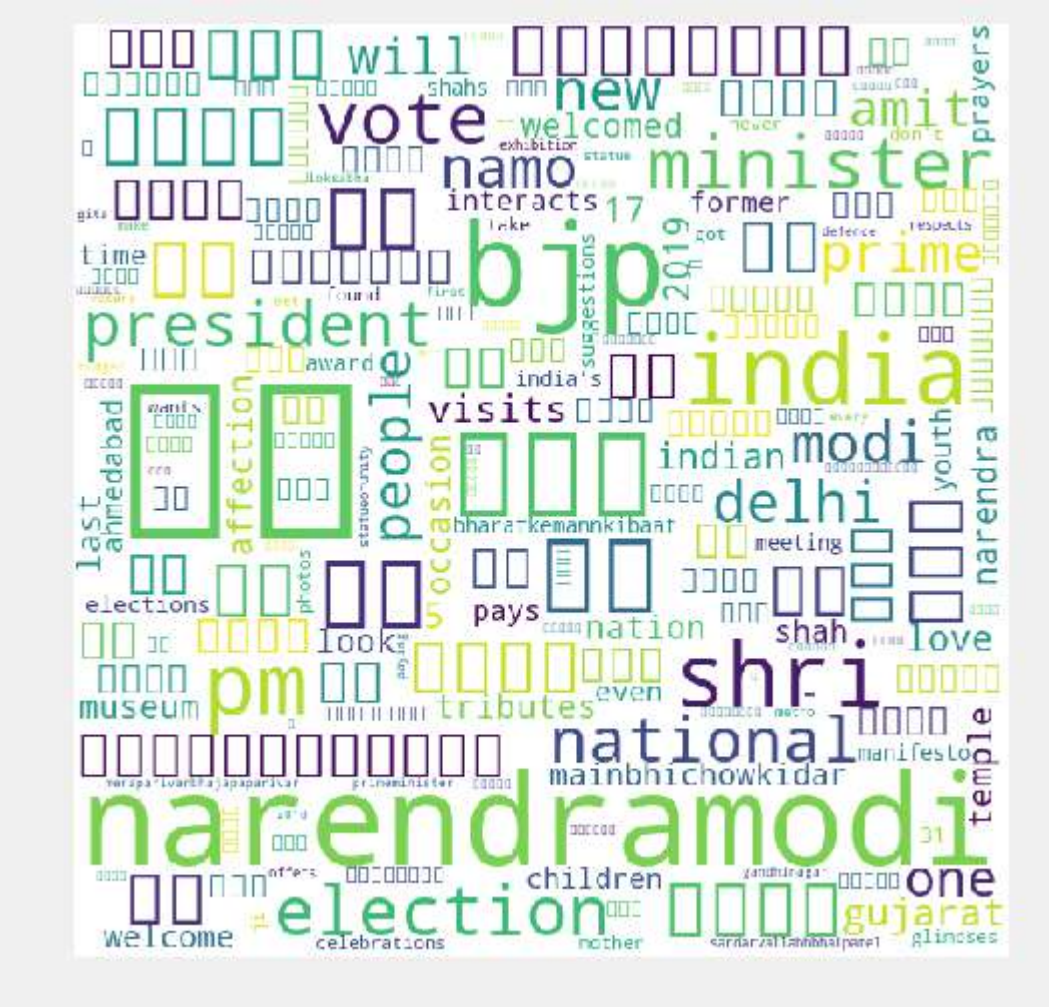
Q2-a Pie chart of share of images and videos in posts

```
value, counts = np.unique([post.is_video for post in posts],
                          return_counts=True)
value = ['Images', 'Videos']
fig1, ax1 = plt.subplots()
ax1.pie(counts, labels=value, autopct='%1.1f%%', shadow=True,
startangle=90)
ax1.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a
circle.
plt.legend(loc="best")
plt.style.use('fivethirtyeight')
plt.show()
```



Q2-b Word cloud of 200 most popular posts

```
# most_popular_posts = sorted(
#     posts, key=lambda x: popularity_of_post(x), reverse=True)[:200]
corpus = reduce(lambda a, b: a + b,
                [post.caption for post in most_popular_posts]).lower()
for c in [
        '.', ',', '"', "'", '?', '!', '/', '\\', ':', ';', '(', ')', '[',
']',
        '{', '}', '&', '@', '#'
]:
    corpus = corpus.replace(c, "")
# corpus = corpus.decode("utf-8")
corpus = corpus.split()
stopwords = set(STOPWORDS)
corpus = [i for i in corpus if i not in stopwords]
wordcloud = WordCloud(
    background_color="white", width=600, height=600, stopwords=stopwords)
wordcloud_dict = Counter(corpus)
wordcloud.generate_from_frequencies(wordcloud_dict)
plt.figure(figsize=(15, 8))
plt.imshow(wordcloud)
plt.axis("off")
plt.show()
```



Q2-c Most used hashtags by BJP's handle

```
hashtags = reduce(lambda a, b: a + b,
                  [post.caption_hashtags for post in posts])
hashtags, count = np.unique([i for i in hashtags], return_counts=True)
hashtags = sorted(zip(hashtags, count), key=lambda x: x[1], reverse=True)
most_popular_hashtags = [hashtag[0] for hashtag in hashtags[:5]]
pprint(most_popular_hashtags)
```

```
['bjp', 'india', 'election', 'vote', 'bharatkemannkibaat']
```

Q3 Collect 3,000 posts from Instagram's 'explore' feed.

```
L.login('username', 'passwword')
posts = []
for post in L.get_explore_posts():
    posts.append(post)
    if (len(posts) == 3000):
        break
```

```
Too many queries in the last time. Need to wait 606 seconds, until 18:34.

Too many queries in the last time. Need to wait 591 seconds, until 18:45.

Too many queries in the last time. Need to wait 586 seconds, until 18:56.
```

```
with open('Instagram_explore_300.pickle', 'wb') as output_file:
    pickle.dump(posts, output_file)
```

```
with open('Instagram_explore_300.pickle', 'rb') as input_file:
    posts = pickle.load(input_file)
```

```
def convert_to_dict(post):
    try:
        x = {
            'caption': post.caption,
            'hashtag': post.caption_hashtags,
            'likes': post.likes,
            'shortcode': post.shortcode,
            'video': post.is_video,
            'type': post.typename
        }
        return x
    except:
        return None
```

```
explore_data = [convert_to_dict(post) for post in posts]
explore_data = [i for i in explore_data if i is not None]
explore_df = pd.DataFrame.from_records(explore_data)
```

Question 3-a Most popular posts

```
explore_df = explore_df.sort_values(['likes'], ascending=False)

display(explore_df[:7])
```

```
.dataframe tbody tr th {
    vertical-align: top;
}
```

```
.dataframe thead th {
    text-align: right;
}
```

	caption	hashtag	likes	shortcode	type	video
1730	None	[]	11223272	Bv4q6GPAeML	GraphImage	False
1445	Spring Break is over 😭 `📖🛍️🍌	[]	6038301	BwADsb1HLDV	GraphImage	False
1231	Sunlight falls into the Abyss \nJust like i fa...	[]	4652382	BwAXtOeHFms	GraphImage	False
916	Coachella better be this good this year @kenda...	[]	3628186	BwInNn6nORx	GraphImage	False
2793	the wind..is this necessary 🦋🦋	[]	2706307	BwDdDRNH4ah	GraphVideo	True
1037	Soho NYC @calvinklein	[]	2654339	Bv9eNr5Do19	GraphImage	False
1361	I love you so much. Thank you Manchester 🌸💙	[]	2267219	Bv_ZTvxnHJt	GraphSidecar	False

selenagomez • Follow

itsbennyblanco 💜💜💜
1w 18439 likes Reply
View replies (327)

juliamichaels 👏👏👏👏👏👏
1w 18206 likes Reply

Liked by joshichetan1996 and 11,309,488 others
7 DAYS AGO

Add a comment... Post

kimkardashian • Follow

kimkardashian Spring Break is over 😭📖📚✏️
5d

kendalljenner P is such a little beachy lady 🥰
5d 41644 likes Reply
View replies (360)

katyperry thank the Lordt
5d 4334 likes Reply
View replies (79)

Liked by mahimaupreti96 and 6,086,380 others
5 DAYS AGO

Add a comment... Post

justinbieber • Follow

justinbieber Coachella better be this good this year @kendalljenner

1d

kendalljenner HAHA i'm dead 😂😂

1d 51455 likes Reply

View replies (321)

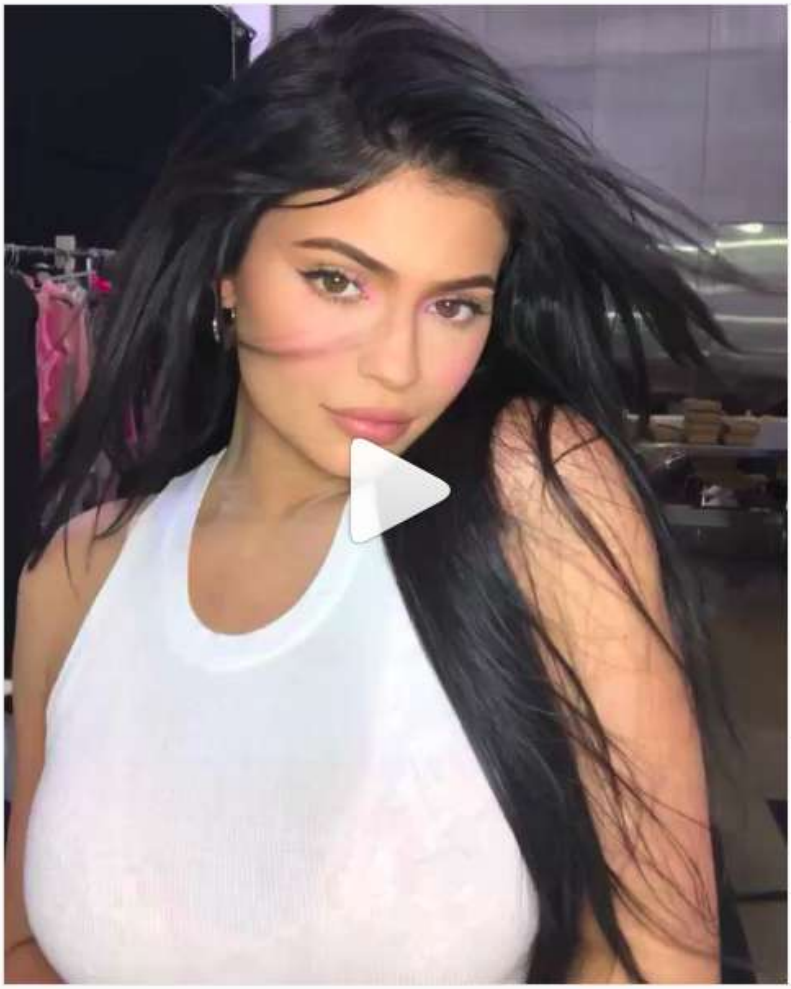
haileybieber hahahahahaha this pic gets me every time

1d 39056 likes Reply

Liked by elliavrram and 4,462,225 others

1 DAY AGO

Add a comment... Post



kyliejenner • Follow

kyliejenner the wind..is this necessary 🦋🦋✨

3d

itsjohnnycyrus INVENTED WIND BLOWING

3d 2226 likes Reply

View replies (340)

afroboss_skigo @temaribb

12h 1839 likes Reply

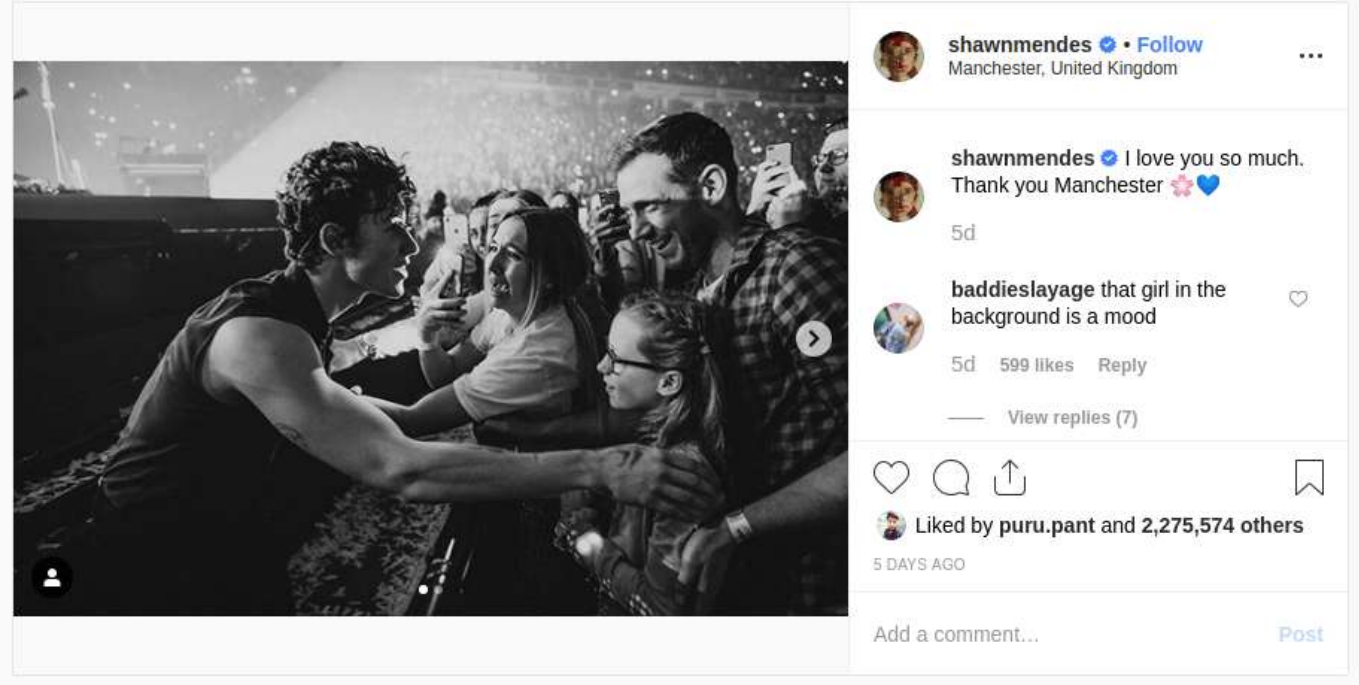
Hide replies

newyorkchity @temaribb 🔥

8,017,457 views

3 DAYS AGO

Add a comment... Post

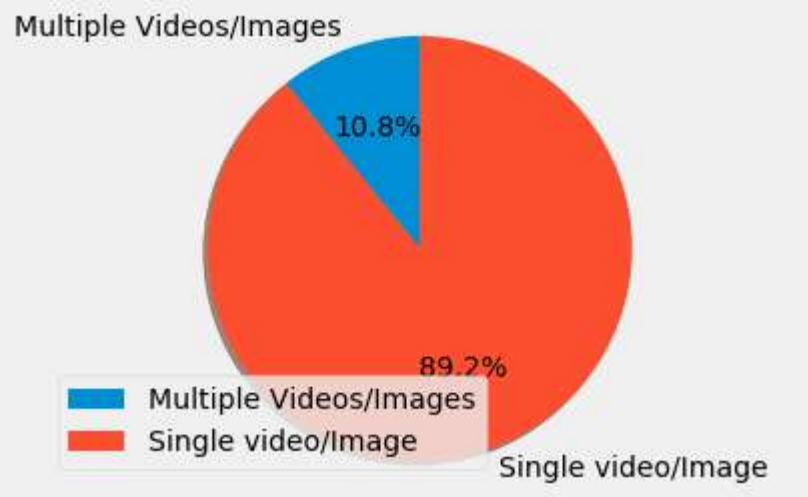


Q3-b Plot pie chart comparing number of posts with single video/ image in them with number of posts with multiple video and images

```
explore_df['type'] = [
    'multiple' if x == 'GraphSidecar' else 'single' for x in
explore_df['type']
]
```

```
value, counts = np.unique(explore_df['type'], return_counts=True)
print(value, counts)
value = ['Multiple Videos/Images', 'Single video/Image']
fig1, ax1 = plt.subplots()
ax1.pie(counts, labels=value, autopct='%1.1f%%', shadow=True,
startangle=90)
ax1.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a
circle.
plt.legend(loc="best")
plt.style.use('fivethirtyeight')
plt.show()
```

```
['multiple' 'single'] [ 325 2675]
```


Multiple Videos/Images
10.8%
89.2%
Multiple Videos/Images
Single video/Image
Single video/Image