Text Analytics Assignment

Sayantan Jana

PDGM-RBA (Finance)

Roll – 42

Topic – Xbox Series X

- The Xbox is a video gaming brand created by Microsoft in the year 2001
- The two most popular gaming console brands are Xbox and PS.
- The new Xbox Series X version is a true rival to the new PlayStation 5.

Importing necessary libraries:

```
import os
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import itertools
import collections

import tweepy as tw
import nltk
from nltk.corpus import stopwords
import re
import networkx

import warnings
warnings.filterwarnings("ignore")

sns.set(font_scale=1.5)
sns.set_style("whitegrid")
```

API keys Authentication:

```
consumer_key= 'c7u5EE29ICqnns6NC0C'
consumer_secret= 'vDt174T8SfDmXarcRXlDIe7MYPLSforfiXTvY98pT'
access_token= '2701903-0Mcuhi5WjSeGZZ7CY7pnavSXdi6BpM52'
access_token_secret= 'mD1hceq9FN8VIV2tbBFvQ9wd1Hq8Hb2di2xSo'

auth = tw.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tw.API(auth, wait_on_rate_limit=True)
```

Search Tweets using keywords:

```
search_term = "#xbox+series+x -filter:retweets"

tweets = tw.Cursor(api.search,
                   q=search_term,
                   lang="en",
                   since='2020-01-01').items(100)

all_tweets = [tweet.text for tweet in tweets]

all_tweets[:5]
```

```
['UK Retailer Warns Of More Potential Xbox Series X Delivery Delays https://t.co/R5MRbC6bwd #Repost #Xbox… https://t.co/kXa8HZWOFx',
 'The Xbox Series X Looks Surprisingly Small Next To The PS5 https://t.co/qj8FGBDmPr #Xbox #XboxSeriesX #PS5 https://t.co/2gJl65sWHt',
 "Please Retweet.\n\nWin a PlayStation 5 or Xbox Series X or $500 USD PayPal #Cash or a $500 USD Gift Card -- Winner's… https://t.co/yMphx3GfIm",
 'You Can Uninstall Different Parts Of Some Games On Xbox Series X|S https://t.co/rrlnhHkSGu #Repost #Xbox… https://t.co/qlggCfQOl7',
 'Walmart\n\nXbox X -  https://t.co/BEqEtchhI3\n\nXbox s - https://t.co/NCopVbn2lQ\n\n#Xbox #launch #BwcDeals #Walmart https://t.co/O4sqGQ1qIa']
```

- Three keywords are used here i.e. Xbox, series and X

Creating a function to remove the URL from tweets:

```
def remove_url(txt):
    """Replace URLs found in a text string with nothing
    (i.e. it will remove the URL from the string).

    Parameters
    ----------
    txt : string
        A text string that you want to parse and remove urls.

    Returns
    -------
    The same txt string with url's removed.
    """

    return " ".join(re.sub("([^0-9A-Za-z \t])|(\w+:\/\/\S+)", "", txt).split())
```

- Python regular expression is used to filter out the https links from the tweets as they are not needed

Cleaned list of tweets:

```
all_tweets_no_urls = [remove_url(tweet) for tweet in all_tweets]
all_tweets_no_urls[:5]
```

```
['UK Retailer Warns Of More Potential Xbox Series X Delivery Delays Repost Xbox',
 'The Xbox Series X Looks Surprisingly Small Next To The PS5 Xbox XboxSeriesX PS5',
 'Please RetweetWin a PlayStation 5 or Xbox Series X or 500 USD PayPal Cash or a 500 USD Gift Card Winners',
 'You Can Uninstall Different Parts Of Some Games On Xbox Series XS Repost Xbox',
 'WalmartXbox X Xbox s Xbox launch BwcDeals Walmart']
```

- Now the tweets are cleaned and stored as a string separated by comma in a python list

Splitting the words in tweets and store as a list:

```
# Create a list of lists containing lowercase words for each tweet
words_in_tweet = [tweet.lower().split() for tweet in all_tweets_no_urls]
words_in_tweet[:2]
```

```
[['uk',
  'retailer',
  'warns',
  'of',
  'more',
  'potential',
  'xbox',
  'series',
  'x',
  'delivery',
  'delays',
  'repost',
  'xbox'],
```

- The tweets are broken into words to count their frequency.
- They are stored as a list inside list

Counting the occurrence of words:

```
# List of all words across tweets
all_words_no_urls = list(itertools.chain(*words_in_tweet))

# Create counter
counts_no_urls = collections.Counter(all_words_no_urls)

counts_no_urls.most_common(15)
```

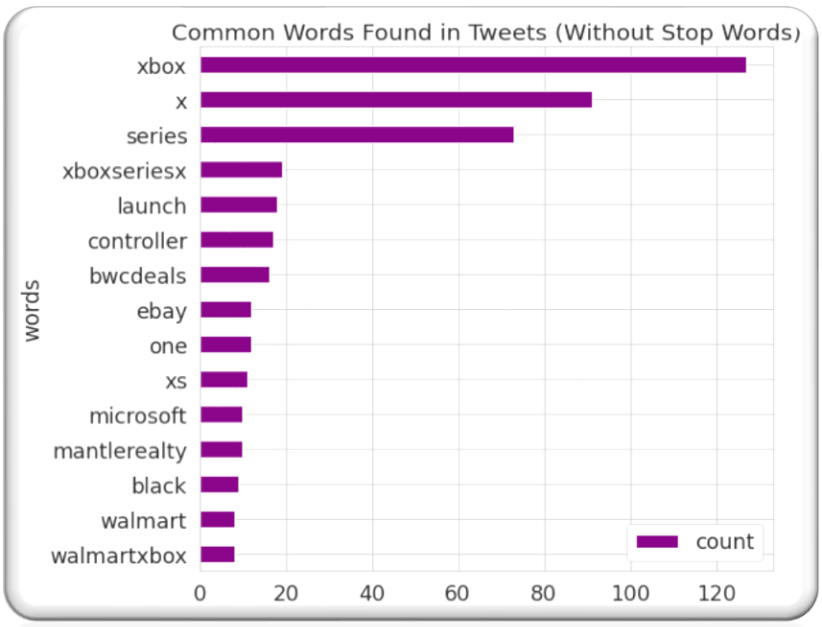
```
[('xbox', 127),
 ('x', 91),
 ('series', 73),
 ('the', 44),
 ('xboxseriesx', 19),
 ('to', 18),
 ('launch', 18),
 ('i', 18),
 ('of', 17),
 ('controller', 17),
 ('for', 17),
 ('bwcdeals', 16),
 ('and', 15),
 ('are', 14),
 ('is', 13)]
```

- The frequency of every word is generated. But there may a bunch of stopwords which is not needed. So they are removed using the following code.

Removing the stop words from the list of words:

```
nltk.download('stopwords')
stop_words = set(stopwords.words('english'))
tweets_nsw = [[word for word in tweet_words if not word in stop_words]
              for tweet_words in words_in_tweet]

tweets_nsw[0]

all_words_nsw = list(itertools.chain(*tweets_nsw))

counts_nsw = collections.Counter(all_words_nsw)

counts_nsw.most_common(15)

clean_tweets_nsw = pd.DataFrame(counts_nsw.most_common(15),
                                columns=['words', 'count'])

fig, ax = plt.subplots(figsize=(8, 8))

# Plot horizontal bar graph
clean_tweets_nsw.sort_values(by='count').plot.barh(x='words',
                      y='count',
                      ax=ax,
                      color="purple")

ax.set_title("Common Words Found in Tweets (Without Stop Words)")

plt.show()
```

Words frequency in a bar graph:



Generating a wordcloud:

```
from wordcloud import WordCloud, STOPWORDS
```

```
def plot_cloud(wordcloud):
    # Set figure size
    plt.figure(figsize=(40, 30))
    # Display image
    plt.imshow(wordcloud)
    # No axis details
    plt.axis("off");
```

```
# Generate word cloud
wordcloud = WordCloud(width = 1368, height = 768, random_state=1, background_color='aquamarine', colormap='Dark2', collocations=False, stopwords = STOPWORDS).generate(for_wordcloud)
# Plot
plot_cloud(wordcloud)
```

- This wordcloud tells what are the words that is most used regarding to the particular tweet, their frequency is directly proportional to their size in the wordcloud.
- The words ‘Launch’, ‘Walmart’, ‘Preorder’, ‘ebay’, ‘bestbuybox’ are some of the most used words.

Get tweets for polarity analysis:

```
from textblob import TextBlob

search_term = "#xbox+series+x -filter:retweets"

tweets = tw.Cursor(api.search,
                   q=search_term,
                   lang="en",
                   since='2020-06-01').items(1500)

# Remove URLs and create textblob object for each tweet
all_tweets_no_urls = [TextBlob(remove_url(tweet.text)) for tweet in tweets]

all_tweets_no_urls[:5]
```

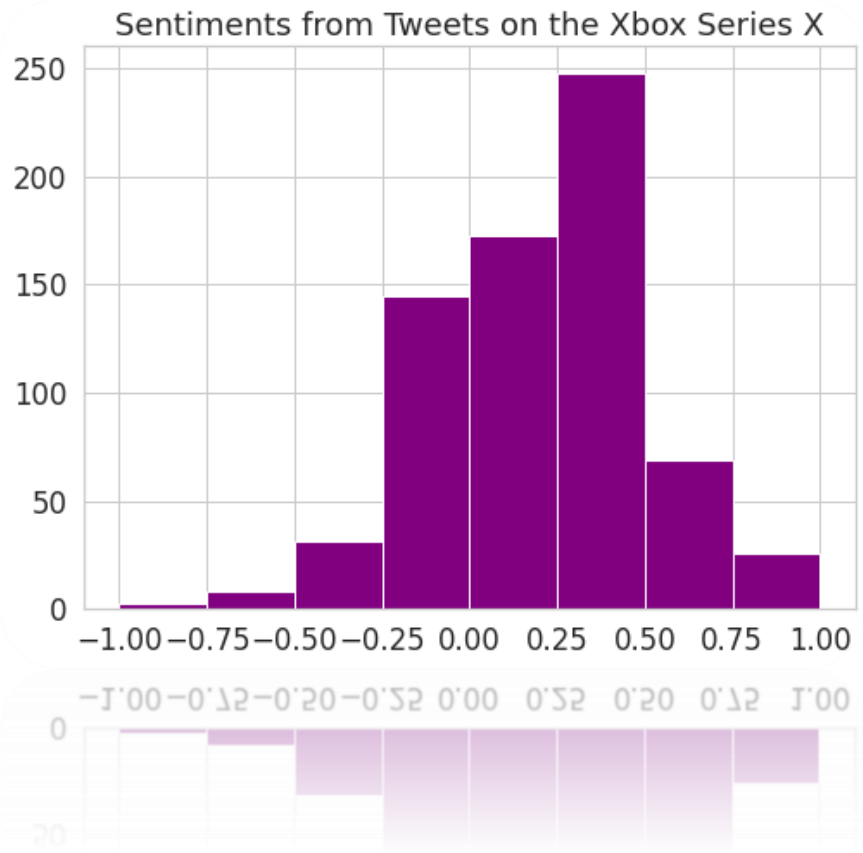
```
[TextBlob("UK Retailer Warns Of More Potential Xbox Series X Delivery Delays Repost Xbox"),
 TextBlob("The Xbox Series X Looks Surprisingly Small Next To The PS5 Xbox XboxSeriesX PS5"),
 TextBlob("Please RetweetWin a PlayStation 5 or Xbox Series X or 500 USD PayPal Cash or a 500 USD Gift Card Winners"),
 TextBlob("You Can Uninstall Different Parts Of Some Games On Xbox Series XS Repost Xbox"),
 TextBlob("WalmartXbox X Xbox s Xbox launch BwcDeals Walmart")]
```

Generate the polarity of individual tweets:

```
# Calculate polarity of tweets
wild_sent_values = [[tweet.sentiment.polarity, str(tweet)] for tweet in all_tweets_no_urls]

# Create dataframe containing polarity values and tweet text
wild_sent_df = pd.DataFrame(wild_sent_values, columns=["polarity", "tweet"])
wild_sent_df = wild_sent_df[wild_sent_df.polarity != 0]

wild_sent_df.head()
```

	polarity	tweet
0	0.250000	UK Retailer Warns Of More Potential Xbox Serie...
1	-0.125000	The Xbox Series X Looks Surprisingly Small Nex...
5	-0.166667	BestBuybox X Black Controller Xbox launch BwcD...
7	0.400000	Xbox One Series X CONFIRMED PRE ORDER from Gam...
10	-0.700000	The X Box Series X and the PS5 both look ugly ...

Plot the polarity in a histogram:



- From the histogram it can be said that people have a positive sentiment regarding the launch of the new Xbox console.