**EXPLORATORY DATA ANALYSIS ON TWITTER API DATA USING PYTHON FOR TOP 3 UNIVERSITIES IN NYC**

**Abstract:**

New York city has several good universities, however there are 3 top universities in NYC which garner the maximum attention from domestic and international students. Our aim here is to do an exploratory data analysis on the tweets posted on the original account of these universities. We will understand the kind of impact that these tweets have on the popularity of the 3 universities.

The following 3 universities have been considered for this analysis:

1. Fordham University
2. New York University
3. Columbia University

**Motivational Statement:**

The main purpose of this analysis was to understand the efficiency of social media marketing strategies utilized by these 3 universities on the twitter platform and how it impacts the reputation of the respective colleges.

**Analysis Methods Used:**

The data extraction has been done using the Twitter API, a developer account was created to extract the tweets for the 3 universities. This analysis has been conducted using several python libraries and modules which are as follows:

1. Tweepy
2. Numpy
3. Pandas
4. Matplotlib
5. Wordcloud
6. Vader sentiment

**Data Collection:**

The data was stored in the following formats to analyze it with the help of python scripts and functions. Dataframe

1. CSV file
2. Dictionary

There was a need to convert the unstructured tweet data extract into a more structured form which is why the above formats were used to store the tweets.

**UNSTRUCTURED FORMAT**

Text, letter

Description automatically generated

**DATAFRAME FORMAT**

Table

Description automatically generated

**CSV FILE**

Table

Description automatically generated

**DICTIONARY FORMAT**

Graphical user interface, text, application, email

Description automatically generated

**Tweet Object Features:**

Each extracted tweet is known as a tweet object, and it has a set of features associated with it. The features are as follows:

1. **tweet.user -** gives tweet object
2. **tweet.id -** gives id of the user posting the tweet
3. **tweet.full\_text** - gives the actual content of the tweet
4. **tweet.created\_at -** gives the date time information
5. **tweet.source -** gives the source of origin for tweet
6. **tweet.retweet\_count -** gives the count of number of retweets for a tweet
7. **tweet.favorite\_count -** gives the number of likes given to a tweet

**Analysis Performed:**

The analysis done on the tweets extracted for Fordham University, New York University and Columbia University is as follows:

1. **Finding the 30 most commonly used words on their twitter accounts**

With the help of this we can analyze the words which are most frequently used on their account and determine the latest topic being talked about.

1. **Representing the most common words as a wordcloud representation**

With the help of this visual representation, we can display the top few words which have the most significant impact on the account.

1. **Conducting Sentiment analysis on the tweets for these universities**

With the help of this we can quantify the kind of tweets that are posted on the account for these universities – positive, negative, or neutral.

1. **Finding the followers for these university accounts & related information**

With the help of this we can quantify the number of followers that a college has and the impact that has on the reach and presence of the college.

The follower information was displayed as a set of the following parameters:

1. **follower.id -** gives the id of the follower
2. **follower.name -** gives the name of the follower
3. **follower.screen\_name** - gives the twitter handle name
4. **follower.location -** gives the location of follower
5. **follower.description -** gives the bio information of follower
6. **follower.url -** gives the twitter url for follower
7. **follower.followers\_count -** gives the number of people who follow this person
8. **follower.friends\_count -** gives the number of people whom this person follows
9. **follower.favourites\_count -** gives the number of likes receivd on this person’s post

**Analysis & Observations:**

* Fordham University twitter account has 23.3K tweets
* Columbia University twitter account has 33.4K tweets
* New York University twitter account has 24.6K tweets
* Most commonly used word in Fordham tweets is about Tania Tetlow the new President
* Most commonly used word in NYU tweets is about their 2 schools NYU Tandon and NYU Law
* Most commonly used word in Columbia tweets is Columbia which is the name of the university
* Data has been extracted in 3 formats, out of them the dataframe format is the most useful to perform data analysis on Jupyter notebook such as working on sentiment analysis & wordcloud
* The data format in csv can also be very useful if the analysis is performed using excel
* Columbia University’s twitter account has the maximum number of followers, followed by NYU and then Fordham University
* This count can be based on both the prestige of the college as well as the content posted on the twitter account
* Columbia university tweets receive the maximum retweets and likes

**Business Recommendations:**

* Fordham University and New York University can work on improving the content posted on their twitter account which can attract more followers and thus it would increase their presence
* Columbia University can try to ensure that the tweets on their twitter handle are more expressive since their sentiment score for most tweets is inclined more towards a neutral value

**Conclusion:**

Based on the analysis conducted on the tweets we can conclude that Columbia university has the maximum presence among users followed by NYU and then Fordham University.