Internship Task 1 report

Topic: Sentimental Analysis of Twitter

Group Members: Ankush Japra, Bhavin Darji, Karan Lakharwal

[**https://github.com/ankushjapra/Sentiments**](https://github.com/ankushjapra/Sentiments)

**Dataset:** US Airline dataset downloaded from Kaggle as a sqlite db file

**Database**: Sqllite3 in python

**Pandas**: The database was read into a pandas dataframe

**Analysis**: The dataset consists of 14,485 records, out of which 2334 are positive, 3069 are neutral and 9082 are negative sentiments. It is obvious there is a huge difference in the distribution of samples.  
All airlines have a higher number of negative tweets as compared to positive and neutral tweets.   
An analysis of reasons for negative tweet shows ‘Customer Service Issue’ being the main problem faced by customers and airlines could improve here to increase customer satisfaction.

Wordcloud was implemented to analyze the occurrence of words for each category: positive, negative, and neutral tweets.   
Apart from airline names, words like reply, service, update, offer, beautiful etc. were most used in tweets. Similarly, for negative tweets the most used words were delay, stuck, waiting, harass, reservations, waited, cancelled etc. These words also help us understand why people were not happy with the airlines. For neutral tweets, the most used words were idea, balance, boarding, travel etc. were the most used words.

**Preprocessing**: The dataset with tweet text was cleaned to convert text to lower case, remove numbers, remove any mentions from the tweets, remove website URLs, special characters and white spaces.

**Feature Engineering:** The cleaned tweets were lemmatized using Wordnet Lemmatizer and thenTF-IDF was used to convert the text into features with a max\_features limit of 2000. We also tried to add tweet length as a feature, but did not find any correlation with the sentiment, so dropped the idea to use it as a feature.