Twitter US Airline Sentiment

Name: Neha Parulekar

**Description of the project & Background Research:**

Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state, or the intended emotional. Anyone who travels regularly recognizes that airlines struggle to deliver a consistent, positive customer experience. The scenario we work through is to estimate the sentiment expressed in tweets about major U.S. airlines.

 Through extensive interview and survey work, the American Customer Satisfaction Index (<http://theacsi.org/>) quantifies this impression, however, The immediacy and accessibility of Twitter provides a real-time glimpse into consumer's frustration:



This Twitter data was scraped from February of 2015 and Contributors were asked to first classify positive, negative and neutral tweets, which was followed by categorizing negative reasons (such as “late flight” or “rude service”) . The fields in the Tweets.csv file / Tweets database table are:

* tweet\_id
* airline\_sentiment
* airline\_sentiment\_confidence
* negativereason
* negativereason\_confidence
* airline
* airline\_sentiment\_gold
* name
* negativereason\_gold
* retweet\_count

The data originally comes from Crowdflower’s Data for Everyone library. Its slightly reformatted version can be found on Kaggle.

For the project I plan on conducting exploratory data analysis to understand and get familiar with the data, and then evaluate then text of the tweets to extract some common themes and get a better and deeper understanding of the reasons for both NEGATIVE and POSITIVE sentiment.

I plan on analyzing the following:

* First finding the proportion of tweets with each sentiment and airline using ggplot
* To find the reasons for tweets with Negative sentiment.
* Try to find the location of tweets through Visualization on maps
* Determine word frequency and build word clouds for each sentiment
* Make cluster analysis of words to understand the associations between words

**Algorithms Used:**

I plan on using the following algorithms for Cluster analysis:

* Connectivity based clustering (hierarchical clustering)

### Centroid-based clustering

For sentiment analysis I plan on using R packages:

* Tm
* SnowballC
* wordcloud

**Evaluation of project:**

The project evaluation can be divided in two parts:

1. The first part would be conducting exploratory data analysis to understand and get familiar with the data, and then evaluate then text of the tweets to extract some common themes.
2. The second would be get a better and deeper understanding of the reasons for both NEGATIVE and POSITIVE sentiment.

**Reference:**

<https://en.wikipedia.org/wiki/Sentiment_analysis>

<https://en.wikipedia.org/wiki/Cluster_analysis>

<http://www.inside-r.org/howto/mining-twitter-airline-consumer-sentiment>