Section 2: Detailed Approach

1. Present an outline of how you are going to go about your investigation.

Objective: To perform text mining and sentiment analysis on Twitter Data and to predict how well the negative comments in social media correlate with market share of company (Anthem Inc.).

Tools & Requirements

- ✓ R and R Studio
- ✓ Twitter App Developer Account

Knowledge Required

- ✓ R Programming,
- ✓ Good Understanding of Social Media and Web Technologies

Major Concerns of My Project

- ✓ How to use R tool**?** (or) How to do R Programming**?**
- ✓ How to get the Twitter Data**?**
- ✓ How to do Text Mining and perform Sentiment Analysis?
- ✓ How negative tweets are correlated with daily change in Stock price of company (Anthem Inc.) ?

A brief summary on project:

To begin with you should have a twitter account before creating twitter app developer account. While creating the twitter app developer account, make sure that you have set the **Call Back URL** as **127.0.0.1:4040** which is local host IP Address and TCP port number. If not, you may get some error when you scrape the tweets from twitter. Next, to scrape the tweets from twitter first of all you have to set up the connection. In order to get the connection, you should have your: Key, Secret Key, Your Access Token and Your Secret Access Token. One can get the details from their twitter app development account. Then you have to enter this data in R console (shown in the below detailed explanation). Next is scraping the tweets with #hashtagname from the twitter and store and store it in the library named tweets and these tweets are stored in documents. Next step in text mining (necessary libraries are loaded) and the tweets are copied to a list named 'mylist' by using 'sapply()' method. Further is text cleaning converts the upper case letters in all the tweets into lower case letters, removes all the numbers from all the tweets, removes all the punctuation marks from all the tweets, removes all the stop words from all the tweets, Removes all white spaces, converts the corpus into Plain Text Document. Next is to perform the sentiment analysis on the data I have used some lexicon which differentiates the negative words from the positive words. The lexicon named **Hu Liu Lexicon,** this lexicon is contributed by Hu and Bing Liu been used. These lexicon files are in working directory as .txt files. For performing the sentiment analysis, I have used the **Jeffrey Breen** approach (detailed code is explained in the later part). I have scraped the tweets with @AskAnthem @AnthemInc @ThinkAnthem @CareMoreHealth @Amerigroup @AskBCBS Ga @AnthemBusiness @AskEmpire @empirebcbs #AnthemBCBS #bcbsAnthem. These tweets are converted into texts. Scores are obtained by adding the positive variables and subtracting negative variables. This gives a histogram showing the score versus frequency and says overall score is neutral. Since the result was not optimal I further performed scatter plot, which did not appear to be much of a trend. So, finally from the graph of sentiment over time we can see that overall sentiment varies by date and seems to have trends over time. Then, I performed logistic regression to find the correlation of negative tweets with stock prices (and obtained the correlation value **-0.9421169**). From the final plot of Percent of negative tweets versus daily change in stock price we can see that logistic regression line fitted has a negative slope indicating that as the percent negative tweets increases the stock price decreases (which is negative correlation).

2. What data will you use? What do you expect to find?

Data Collection:
Data is extracted from Twitter API. The tweets are collected using Twitter API and filtered by performing sentiment analysis using R programming.

- ✓ Data for this project will come from two sources - tweets about Anthem and daily stock prices from Yahoo.com. Tweets will be collected using the twitteR package which uses the Twitter API to search and return tweets. Stock prices will be downloaded as a CSV and loaded into R directly.
- ✓ Consumers use many different Twitter accounts and hashtags to reach Anthem. The ones used for this analysis are shown below: @AskAnthem @AnthemInc @ThinkAnthem @CareMoreHealth @Amerigroup @AskBCBSGa @AnthemBusiness @AskEmpire @empirebcbs #AnthemBCBS #bcbsAnthem.
- ✓ Once tweets have been gathered the above accounts and hashtags a sentiment analysis will be performed to determine the number of positive and negative words contained in each tweet. The resulting score will indicate the overall sentiment of the tweet.
- ✓ Once sentiment scores are created a logistic regression analysis will be performed to determine the relative associate between daily sentiment (daily percent of tweets that are positive) with daily change in stock price (opening price - closing price).

Data Pre-Processing: Tweets consists of many acronyms, emoticons and unnecessary data like pictures and URL's. So tweets are preprocessed to represent correct emotions of public. For preprocessing of tweets we employed three stages of filtering: Tokenization and stop words removal for removing special characters.

- ✓ **Tokenization:** Tweets are split into individual words based on the space and irrelevant symbols like emoticons are removed. We form a list of individual words for each tweet.
- ✓ **Stop word Removal:** Words that do not express any emotion are called Stop words. After splitting a tweet, words like a, is, the, with etc. are removed from the list of words.

Sentiment Analysis:
Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service. It generally aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. Tweets are classified as positive, negative and neutral based on the sentiment present.

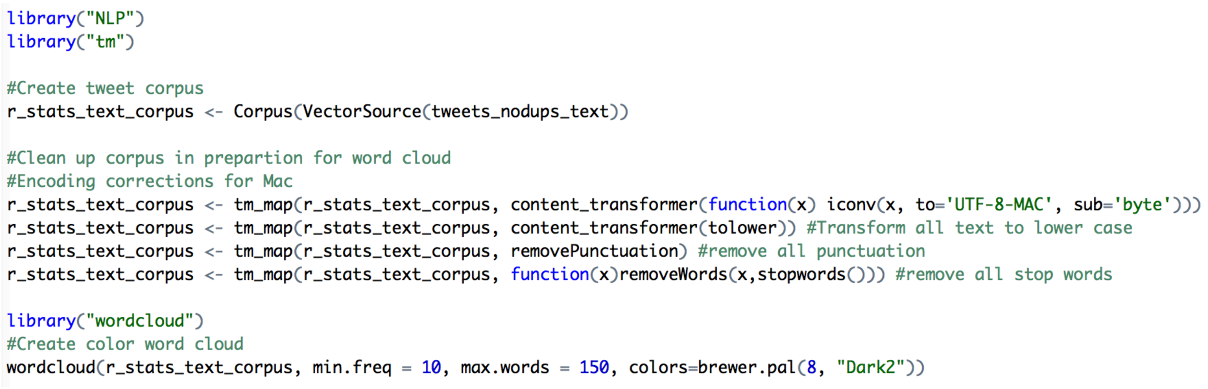
Anthem Inc., is the second largest health insurance provider (by membership) in the United States and the largest Blue Cross Blue Shield member organization. For this project I have performed a sentiment analysis on tweets about Anthem, Inc. and examine the relationship between sentiment and stock prices.

3. Document initial findings, and include descriptive statistics from your data in tabular and graphical formats.

Exploring the tweets:

Now that I have the data in the formats I need it in for analysis I can do some preliminary exploration. First I will examine a word cloud of the tweets to see the types of words that make up the returned tweets (Below is the R code for word cloud).

A tag cloud (word cloud, or weighted list in visual design) is a visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color.

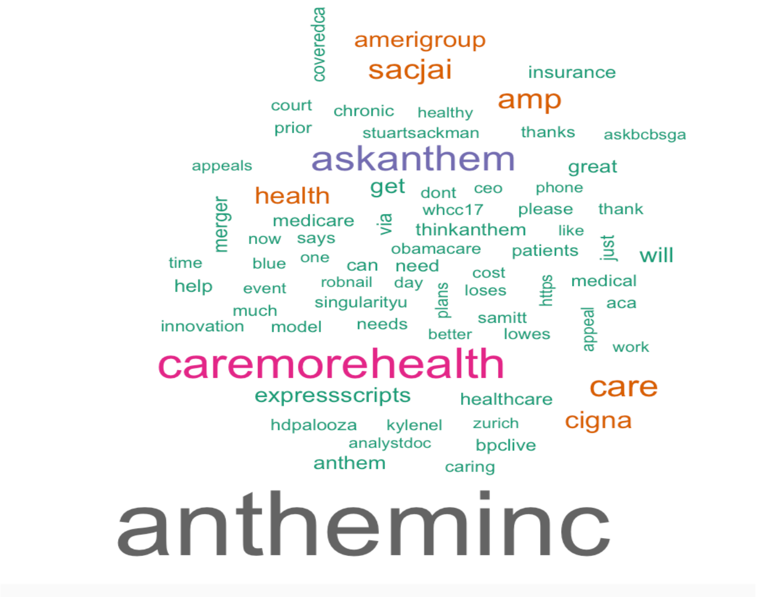


```
library("NLP")
library("tm")

#Create tweet corpus
r_stats_text_corpus <- Corpus(VectorSource(tweets_nodups_text))

#Clean up corpus in prepartion for word cloud
#Encoding corrections for Mac
r_stats_text_corpus <- tm_map(r_stats_text_corpus, content_transformer(function(x) iconv(x, to='UTF-8-MAC', sub='byte')))
r_stats_text_corpus <- tm_map(r_stats_text_corpus, content_transformer(tolower)) #Transform all text to lower case
r_stats_text_corpus <- tm_map(r_stats_text_corpus, removePunctuation) #remove all punctuation
r_stats_text_corpus <- tm_map(r_stats_text_corpus, function(x)removeWords(x,stopwords())) #remove all stop words

library("wordcloud")
#Create color word cloud
wordcloud(r_stats_text_corpus, min.freq = 10, max.words = 150, colors=brewer.pal(8, "Dark2"))
```



Screenshot showing the Word Cloud

The word ‘antheminc” has repeated most times in the tweets as its font size in larger than all. Words with same color and same font size indicates that it has same frequency (number of repetition is same)