STATS133 Final Project

Nick Rodriguez, Navya Putta, Anna Yea Jung, Bret Hart

December 12, 2015

A Sentiment Analysis of Twitter Text

Purpose of the Project

The question we wanted to answer is simple. How do people feel about certain fast food establishments? We attempted to answer this question by taking a look at the tweets that people messaged to certain resturants' Twitter accounts.

Text data has always been a rich source of data, and we feel that analysts should become familiar with working with it.

In order to answer our question, we had to discover how to measure of people's sentiments. The team ended up utilizing useful packages in order to perform the desired analyses.

Data Extraction

The data came from many sources essentially that was pipelined through the **twitteR** package. We collected the tweets that were messaged to a certain restaurant's Twitter account.

Here is an example query that we would send in a Twitter session in R:

```
McDsample <- searchTwitter(searchString = "to:McDonalds" , n = 200, lang = "en")
```

However, to interact with Twitter, you'll need to go through the authentication process. There are directions to do this in the *DataAcquisition.R* file which can be found in the subdirectory, *DataAcquisition*, of the *RawData* directory.

A Side Note

If you are interested in Twitter data, but would like to obtain it by other means then we suggest you do it through Python. UC Berkeley's D-Lab possesses a great walkthrough on how to interact with the Twitter API in Python. However, either method you choose, you will need to set up a Twitter Developer's account.

Cleaning the Raw Data

Our data came in the form of character vectors. With that being said, we worked closely with the regular expressions in order to clean the tweets. We had to remove retweet instances, punctuation and other unneeded strings like emojis. We were able to take advantage of R's tendency to vectorize data, which makes our cleaning script very powerful for Twitter text data.

Transforming the Data Further

After the cleanup, we possessed vectors of character type data. We needed to conduct sentiment analysis to actually answer our question. The package **sentiment** was able to provide us with the methods we needed to get this data. However, the **sentiment** package is no longer supported on CRAN, so you must download the source code. Here are the commands in order to achieve this.

```
install_url("http://cran.r-project.org/src/contrib/Archive/Rstem/Rstem_0.4-1.tar.gz")
install_url("http://cran.r-project.org/src/contrib/Archive/sentiment/sentiment_0.2.tar.gz")
```

Classifying Tweets By Emotion