**Data Science Application:**

Sentiment Analysis and Movie Review

Edaño Jr., Celestino B.

1313915

<www.celcs405.tk>

October 16, 2016

Technological Institute of the Philippines, Quezon City

**Introduction**

The evolution of web technology has led to a huge amount of user-generated content and has significantly changed the way we manage, organize and interact with information. Due to the large amount of user opinions, reviews, comments, feedbacks and suggestions it is essential to explore, analyze and organize the content for efficient decision making. In the past year’s sentiment analysis has emerged as one of the popular techniques for information retrieval and web data analysis. Sentiment analysis, also known as opinion mining is a subfield of Natural Language Processing (NLP) and Computational Linguistics (CL) that defines the area that studies and analyzes people’s opinions, reviews, and sentiments.

Sentiment Analysis has gained much attention in recent years and emerged as a popular and efficient technique for information retrieval and web data analysis. It refers to the use of natural language processing, text analysis, and computational linguistics to identify and extract subjective information in source materials. The exponential growth of the user generated content has opened new horizons for research in the field of sentiment analysis.

According to Fang and Zhang in their paper entitled “Sentiment analysis using product review data”, sentiment is an attitude, thought, or judgment prompted by feeling. From a user’s perspective, people are able to post their own content through various social media, such as forums, micro-blogs, or online social networking sites. From a researcher’s perspective, many social media sites release their application programming interfaces (APIs), prompting data collection and analysis by researchers and developers. For instance, Twitter currently has three different versions of APIs available, namely the REST API, the Search API, and the Streaming API. With the REST API, developers are able to gather status data and user information; the Search API allows developers to query specific Twitter content, whereas the Streaming API is able to collect Twitter content in real-time. Moreover, developers can mix those APIs to create their own applications. Hence, sentiment analysis seems to have a strong fundament with the support of massive online data.

**Background of the Study**

Our ability to derive social and economic value from the newly available data is limited by the lack of expertise. Working with this data requires distinctive new skills and tools. The corpuses are often too voluminous to fit on a single computer, to manipulate with traditional databases or statistical tools, or to represent using standard graphics software. The data is also more heterogeneous than the highly curated data of the past. Digitized text, audio, and visual content, like sensor and blog data, is typically messy, incomplete, and unstructured; it is often of uncertain provenance and quality; and frequently must be combined with other data to be useful. Working with user-generated data sets also raises challenging issues of privacy, security, and ethics.

Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service to identify the feelings of the consumers. In social media, such as product reviews, sentiment analysis can be used to uncover whether consumers are satisfied or dissatisfied with the product and the services offered by the company. Likewise, a company could use sentiment analysis to measure the impact of a new product, ad campaign, or consumer’s response to recent company news on social media.

Sentiment analysis is often used in business intelligence to understand the subjective reasons why consumers are or are not responding to something. For example, why are consumers buying a product? Why consumers still prefer their company service? What do they think of the user experience? Did customer service support meet their expectations? Sentiment analysis can also be used in other areas like political science, sociology, and psychology to analyze trends, ideological bias, opinions, gauge reactions, etc.

This study focuses on making a review of the top three grossing movie films for each month of 2016. The review is based on the sentiment analysis which uses Twitter’s different versions of API’s, machine learning, and natural language processing approach. The key challenges for sentiment analysis that I assume to encounter while working on this project are: *named entity recognition*, which refers to the person actually talking about, e.g. is 300 Spartans, does it refer to a group of Greeks or a movie. Second is *anaphora resolution* - the problem of resolving what a pronoun or a noun phrase refers to. "We watched the movie and went to dinner; it was awful." What does "It" refer to? Next is *parsing* - what is the subject and object of the sentence, which one does the verb and/or adjective actually refer to? Then we have *sarcasm* - if you don't know the author or the author’s intention or style you have no idea whether bad means bad or good. And lastly the *Twitter* itself, which is full of abbreviations, lack of capitals, poor spelling, poor punctuation, poor grammar, and much more.

**Data Structures and Knowledge Representation**

The data structures and knowledge representation that I was used are the following:

* **Data Frame** -  is a table, or two-dimensional array-like structure, in which each column contains measurements on one variable, and each row contains one case. I used this to store the polarity of each tweets.
* **Sentiment package in R** – a special package that classifies the emotion of the text.
* **Word cloud** - a handy tool used to highlight the most commonly cited words in a text using a quick visualization.
* **Ggplot2 package in R** - create graphs that represent both univariate and multivariate numerical and categorical data in a straightforward manner. Grouping can be represented by color, symbol, size, and transparency. The creation of trellis plots (i.e., conditioning) is relatively simple.

**Software and Programming Languages**

*Software*

The following are the software that I will be going to use:

* R-Studio (any version) – this software is needed to perform sentiment analysis. It acts as a user interface for R Programming Language.
* Sublime Text 2 – used to develop a simple website. It’s a sophisticated text editor for code, markup, and prose.
* Hostinger – this is where I created the domain name for my website and uploaded my HTML files.
* Freenom – used to run my website.

*Programming Language*

The following are the programming languages that I will go to use:

* R Programming Language – use to perform sentiment analysis through Twitter API.
* HTML, CSS, JavaScript, and JQuery – create a simple website for reviews.

**Approaches and Algorithms**

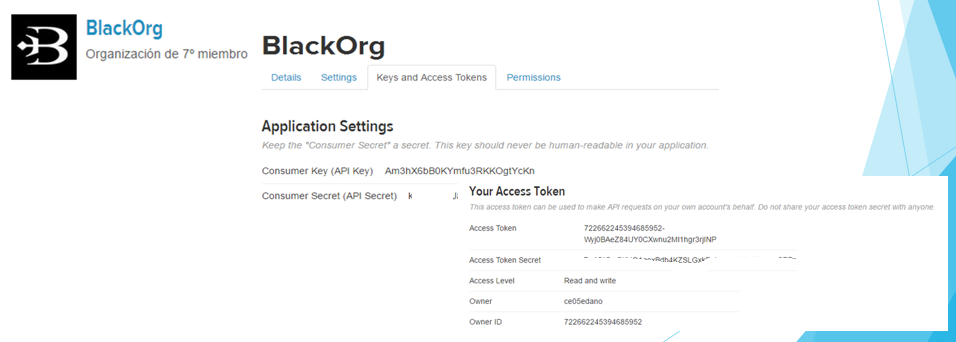
The following are the Natural Language Processing technologies that I will go to use on my project:

* **Sentiment Analysis** - the use of natural language processing, text analysis, and computational linguistics to identify and extract subjective information in source materials.
* **Information Extraction** - extracting structured information from unstructured or semi-structured machine-readable documents. In most of the cases, this activity concerns processing human language texts by means of natural language processing (NLP).
* **Twitter Data Analysis** – focuses on data collecting and data mining on Twitter.
* **Naïve Bayes algorithm** - a Machine Learning technique to characterize subjective human opinions or sentiments.
* **Reviews (** [**www.celcs405.tk**](http://www.celcs405.tk) **)** – making a review based on the collected information and sentiment analysis.

**Experiment and Results**

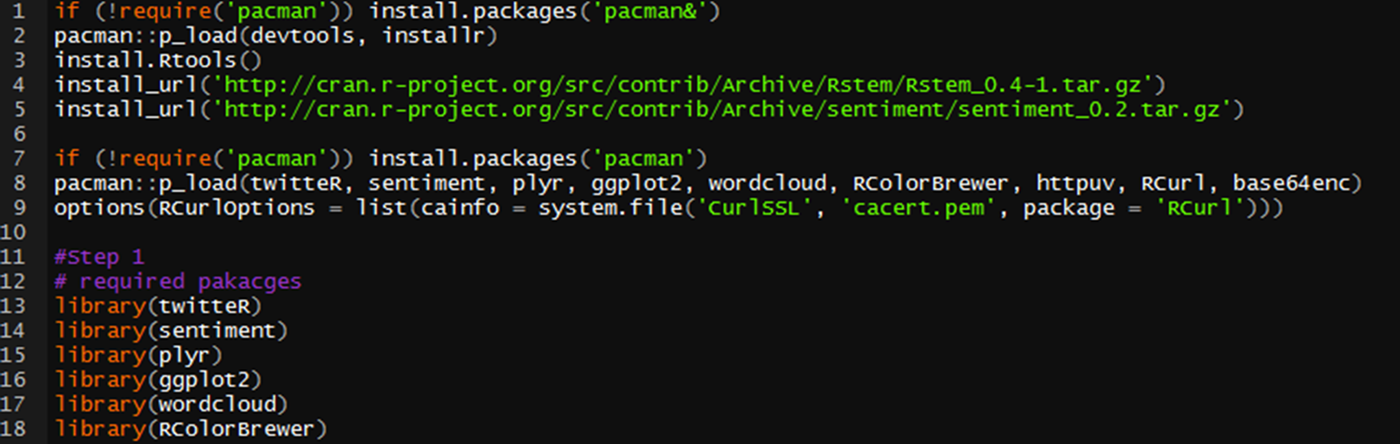
The following are the steps I used to gather tweets from Twitter, cleanse each tweets, perform sentiment analysis and results:

***Step 1:*** *Creating an application at apps.twitter.com and generate API key, Access tokens, etc.*



**Figure 1 shows how to generate API keys and tokens.**

***Step 2:*** *Authorize R to access Twitter.*

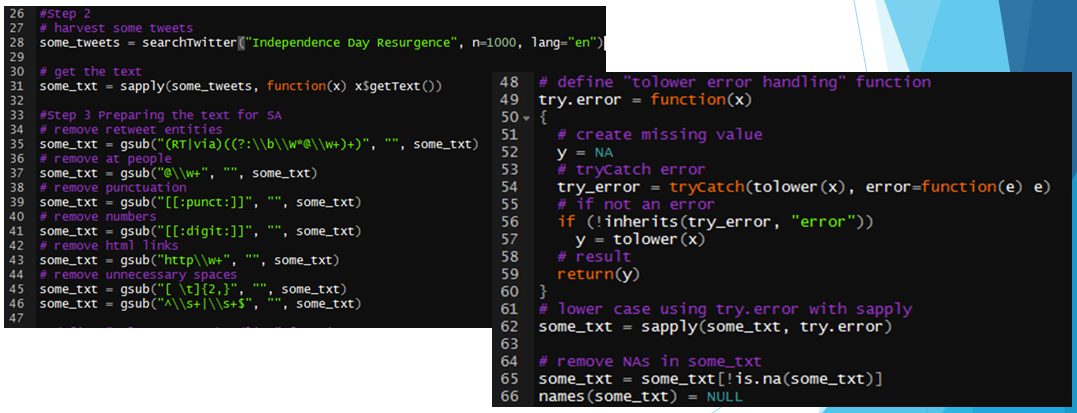


**Figure 2 shows how to download all the necessary libraries.**

The following are the R package/libraries that I used.

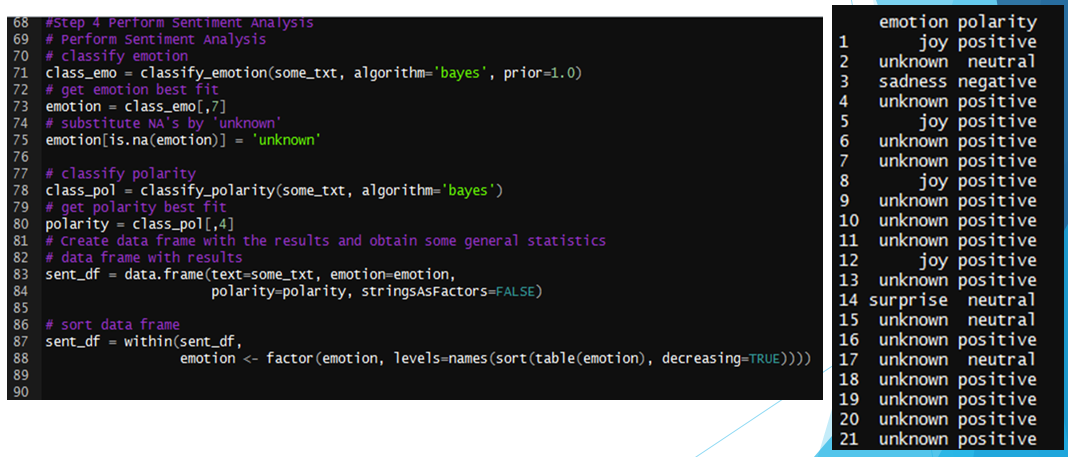
* **twitteR:** which gives an R interface to the Twitter API.
* **sentiment:** classifies the emotions of text.
* **plyr:** for splitting text.
* **ggplot2:** for plots of the categorized results.
* **wordcloud:** creates word clouds of the results.
* **RColorBrewer:** color schemes for the plots and word cloud.
* **httpuv:** required for the alternative web authorization process.
* **RCurl:** http requests and processing the results returned by a web server.

***Step 3:*** *Perform a search and do text normalization for each tweets.*



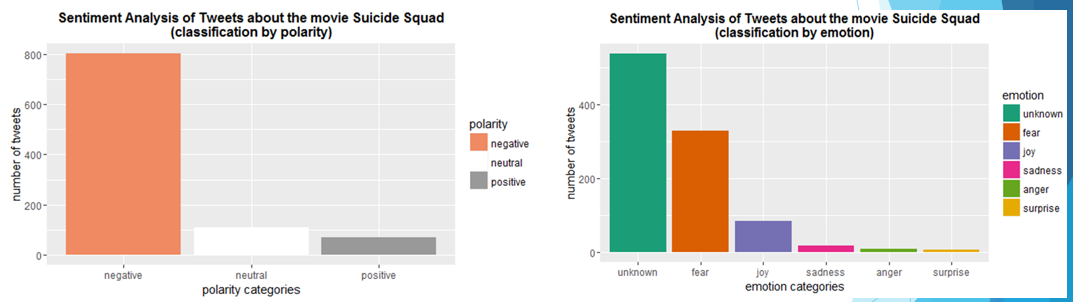
**Figure 3 shows how to clean each tweets and deal all those out of vocabulary words.**

***Step 4****: Perform sentiment analysis*



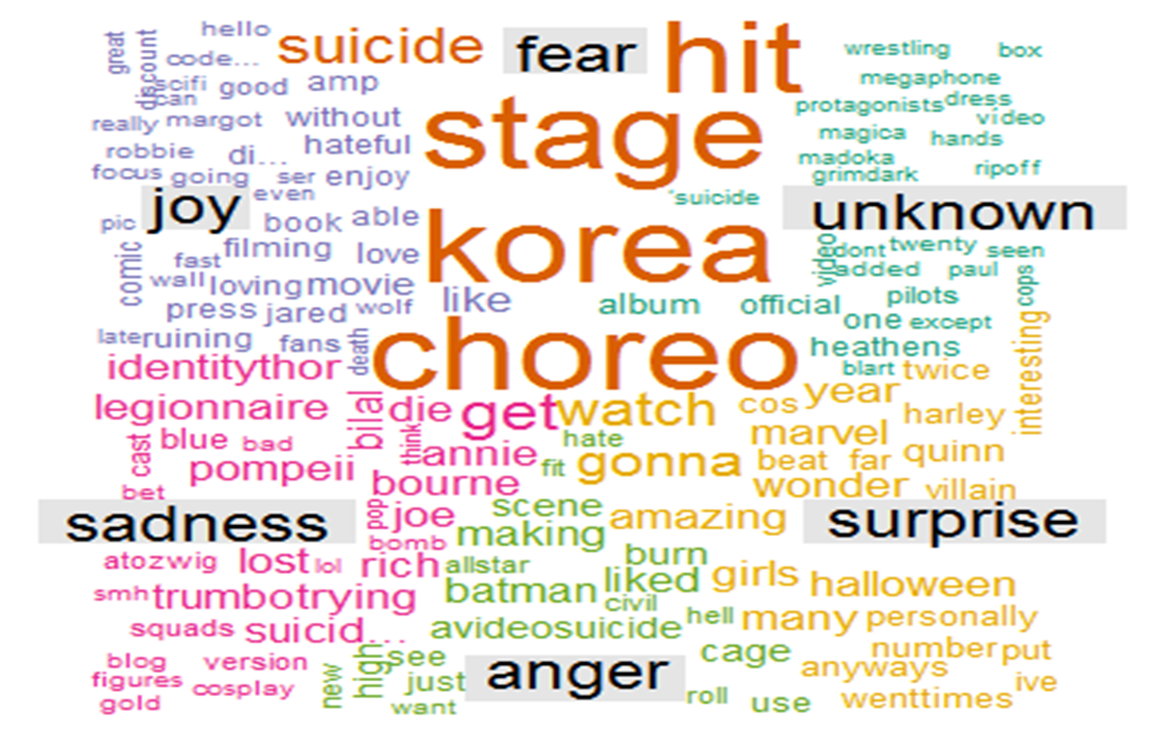
**Figure 4 shows how to apply Naïve Bayes algorithm to define the polarity of the tweet.**

***Step 5:*** *Plotting the obtained results based on polarity and emotions.*



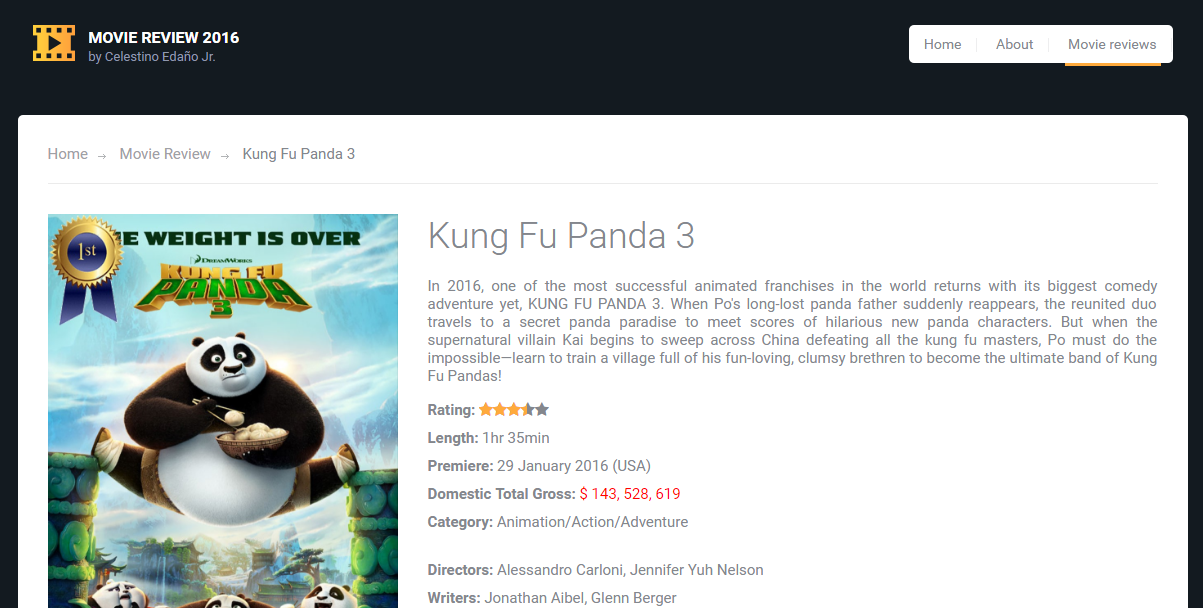
**Figure 5 shows the plot of obtained results based on polarity and emotion of each tweets.**

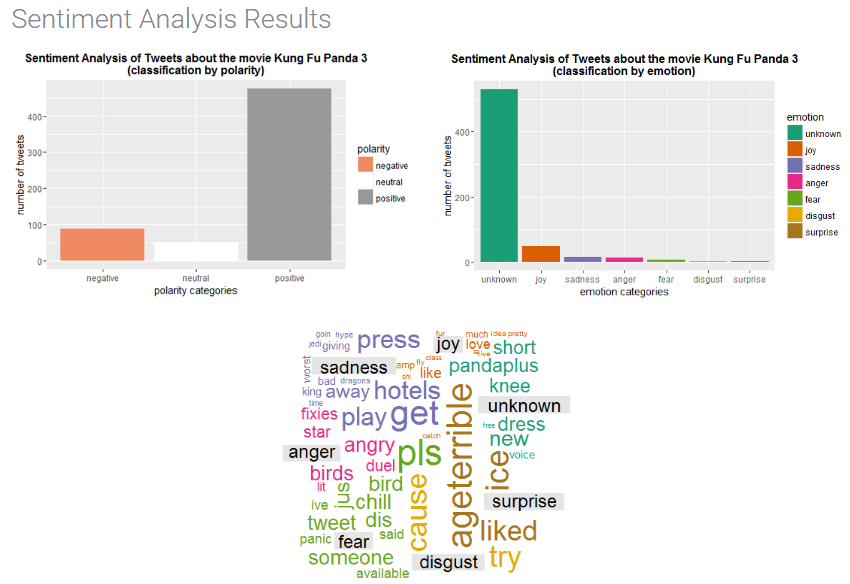
***Step 6:*** *Creating the word cloud.*



**Figure 6 shows the word cloud generated.**

***Step 7:*** *Create a simple website that contains all necessary information. (* [*www.celcs405.tk*](http://www.celcs405.tk)*)*





**Figure 7 shows the simple website that contains all the information based on the conducted sentiment analysis.**

**Relationship to Artificial Intelligence and Machine Learning**

My project sentiment analysis or opinion mining is related to Artificial Intelligence because opinion mining is one of the task in Natural Language Processing which is a field of Artificial Intelligence. It is related to Machine Learning because I used Naïve Bayes algorithm which is a Machine Learning technique. And over the year’s machine learning makes sentiment analysis more convenient. My project introduced how to do sentiment analysis with machine learning using R. In the landscape of R, the sentiment R package.

**Conclusion**

Data is increasingly cheap and ubiquitous. We are now digitizing analog content that was created over centuries and collecting myriad new types of data from web logs, mobile devices, sensors, instruments, and transactions. At the same time, new technologies are emerging to organize and make sense of this avalanche of data. We can now identify patterns and regularities in data of all sorts that allow us to advance scholarship, improve the human condition, and create commercial and social value. The rise of "big data" has the potential to deepen our understanding of phenomena ranging from physical and biological systems to human social and economic behavior. The field of data science is emerging at the intersection of the fields of social science and statistics, information and computer science, and design.

Within past decade Artificial Intelligence’s different field such as Natural Language Processing, Expert Systems, Neural Network, etc. has been the spotlight for the past 40 years or more. With that said, performing sentiment analysis to make a movie review is much easier. With the help of some machine learning technique like Naïve Bayes, Support Vector Machines, Random Forest, Perceptron and Neural Networks etc. doing sentiment analysis is more convenient than before.