**Data Cleaning Process**

Raw data is the original data and is often hard to use for analysis directly. Data processed after applying the data cleaning step is ready for analysis. For this reason data cleaning process plays a very crucial in our Project. The data we extracted from twitter contains many undesirable symbols, words, and punctuations etc. that may slow down our applied algorithms or even make them to perform incorrectly. Hence, it becomes very necessary to write data cleaning modules or in some cases use the language library for the same purpose.

We present some of the differences between the rawdata and the tidy data obtained after the cleaning module is applied on it:

* Raw data may only be need to process only once (during cleaning).
* Raw data is the original data and is often hard to use for analysis directly.
* Processed Data is ready for analysis.
* Processing of raw data includes removal of hashtags, links, usernames, punctuations, non-English words and symbols like RT.

Pre-processing of Tweets

To prepare the transaction file that contains opinion indicators, namely the adjective, adverb and verb .

Thus, we pre-process all the tweets as follows:

1. Remove all URLs (e.g. www.example.com), hash tags (e.g. #topic), targets (@username), special Twitter words (“e.g. RT”).
2. Calculate the percentage of the tweet in Caps.
3. Correct spellings: A sequence of repeated characters is tagged by a weight. We do this to differentiate between the regular usage and emphasized usage of a word.
4. Using a POS tagger, the NL Processor linguistic Parser, we tag the adjectives, verbs and adverbs.

The following strategy has been followed to make the downloaded twitter data tidy

1. For Removal of hashtags, a substring removal python function is used.
2. Similar to the removal of hashtags, substring removal function is used to remove the links in the tweets.
3. Each punctuation is replaced by an empty substring
4. Removal of strings generally present in tweets such as RT, @someUser etc. through the substring removal function.
5. Removal of non-English tweets by encoding them into ASCII and if encoding produces any exception, then reject them, otherwise, select the tweets.
6. Removal of stopping words through nltk library.
7. The substring removal function removes the part of the string that contains a substring e.g. if substring = 'http', then http://www.google.com is removed, that means, remove until a space is found.
8. Substring removal function has been very useful in removing the links, as well as retweets, hashtags and some other undesired content from the tweets.
9. NLTK is a leading platform for building Python programs to work with human language data.
10. It provides easy-to-use suite of text processing libraries for tokenization and tagging.
11. NLTK has been used to remove the stopping words from all the tweets.
12. Words are fed to the Part of Speech tagger after Sentence Tokenizer and Word Tokenizer is applied.
13. Only adjectives, adverbs and verbs have been retained.All other words are discarded.