ASSIGNMENT C4

Title: Twitter Data Analysis

Problem statement:

Use Twitter data for sentiment analysis. The dataset is 3MB in size and has 31,962 tweets. Identify the tweets which are hate tweets and which are not.

Objective:

- Perform twitter data sentiment analysis
- Understant basic NLP and text feature extraction.

Outcomes:

We will be able to:

- Learn steps to perform sentiment analysis
- Learn text feature exactions
- Learn NLP concepts

Hardware and software requirements

- OS: Fedora 20 / Ubuntu (64-bit)
- RAM: 4GB
- HDD: 500 GB
- Jupyter Notebook
- Python libraries

Theory:

- Sentiment analysis is a process of determining whether a piece of writing (product/movie review/tweet, etc) is positive, negative or neutral.
- It can be used to identify customer's or follower's attitude toward a brand through the use of a variable such as context, tone, emotion, etc.

→ steps to perform Sentiment Analysis:

1. Gather relavant tweets from Twitter.

2. Preprocessing:
   - It is an essential step to make the raw text ready for mining i.e. it becomes easier to extract information from the text and apply machine learning to it.
   - If we skip this step then ther is a higher chance that you are working with noisy & inconsistent data.

3. Feature extraction:
   - Selection of useful words from the tweets is called as feature extraction. In this method, we extract this aspect from preprocessed dataset.
   1. There are 3 different types of features namely unigram, ~~by~~ bigram and n-gram features.
   2. Parts of speech tag such as like adjectives, verbs, adverbs and nouns are good indicators of subjective sentiment.

3. Negation is another important but difficult feature to interpret. The presence of negation usually changes the polarity of the sentiment.

4. Feature selection:
- Correct feature selection technique are used in sentiment analysis that has got a significant role for identifying relavant attributes and increasing classification accuracy.

5. Classification:
- For classification of tweets & naive bayes algorithm -

a. Naive Bayes is a probabilistic classifier inspired by bayes algorithm under a simple assumption which is attribute are conditionally independent.

$$P(c|x) = \frac{P(x|c)\ P(c)}{P(x)}$$

The classification is conducted by deriving the maximum posterion which is maximal $P(c_i|x)$.

b. SVM Classifier:
- It is known to perform well in sentiment analysis,
- SVM analyses the data, define the decision boundry and uses the kernel for computation which are performed in input state.
- The input data are two sets of vector, each of size m. Then every data which is represented as vector is classified into a class

Conclusion:

Hence, we have successfully predicted tweets to be positive or negative. from the given dataset.