Gathering , Storage and Sentiment Analysis of Geo-Tagged Tweets.

Members

Qais Makani(A029)
Yash Rathod(A042)
Chinmay Parikh(A059)

Abstract

Due to leaps in social media and technology, Almost everyone is connected via a social network. Twitter is the most popular of the social media sites. It uses a simple follow model (you only recieve updates from the people you follow). Users update their wall by posting "tweets"."Tweets" are one short (140 character limit) updates.

> Companies which also have a twitter account are followed by their customers.Such companies desire to know the how well their product is doing on the market by analyzing customer tweets.This analysis for positive and negative feedback is known as "Sentiment Analysis".

Its easy for human to decipher the tweets left by the customers and sort them into the three categories.But computers don't understand human language or the context of the tweet or concepts such as sarcasm. The application of **Natural Language processing** together with **Machine Learning** helps close the gap between computers and humans.

Sentiment Analysis uses ***Machine Learning*** to train a "Classifier". The "Classifier" is trained using training data, In which tweets are fed with "labels" to the "Classifier".

Example of the Training Data set:

```
"I love Holidays" , "Positive"
"I hate when spiders appear out of thin air" , "Negative"
"Choclate is the BEST" , "Positive"
"The Rains are depressing" , "Negative"
```

Then Feature Extraction using ***Natural Language processing*** to extract "Features" from the tweet.

We Get :

```
'Love' , Positive probability 97.43%
'hate' , Negative probability 94.64%
'BEST' , Positive probability 88.41%
'depressing' , Negative probability 84.43%
```

After Training the "Classifier" we can now Use it on test data. One may automate this so as to allow for self-learning.

Goals

Listed According to priority

1. Prototype Which Runs on a pre-compiled "Corpus". Prototype Deliverable.
2. Enhance the Tweet Gathering System to make it Easy to use. GUI Version Deliverable.
3. Implement Either the Naive Bayes Classifier or the Support Vector Machine. Alpha Stage Deliverable.
4. Enhance the Feature Extraction and add Stemming or other methods to improve Accuracy. Alpha Stage Deliverable.
5. Implement a World map to display the real power use of the Geo-tagged property of the tweets. This allows companies to track user's of their product and the regional areas in which their product is facing problems. Beta and Final Product Deliverable.

The Three Main Parts

1. Get The Tweets - Twitter Streaming API.
2. Parse and Store the Tweets - JSON and SQLITE
3. Analyze the Tweets. - Sentiment Analysis.