MOTOR TWEET VOGUE (MTV)
Project CSE 6339

Objective and overview of the project:

The objective of the project is to identify the trends in the car models in each states of the country using Twitter data. This is identified based on the popularity of each car model in each state.

Popularity will be calculated initially based on the number of times each model is mentioned in the twitter data. In a later stage sentiment analysis will be performed on the data and positive opinions will be considered to up vote and negative opinions to down vote the popularity measure. Sentiment Analysis will be performed in two ways, one using the opinion information in the tweet and another using emoticons.

Popular models in each state and popularity of each model in different states over a selected period of time will be presented.

Challenges:

Amount of data: Request Limit of Streaming API is 150 calls/15 min. Hence the amount of Data to be considered has to be limited. Since streaming API returns current data, the amount of tweets that has location information varies. This also affects the amount of tweets that has to be considered for the work.

Location of Tweets: Since location information is not a mandatory field, tweets can or cannot have location details which can be either location geocode or manually provided location information. Genuineness of manually provided location information is not guaranteed, which is another challenge. This may affect the result.

Entity Recognition: Entity recognition is another challenge. Entity Recognizer should be able to distinguish car names from other entities (for e.g. jaguar, Avenger, Rogue.) and Nick names.

Sentiment Analysis: Sentiment analysis is another challenge. Because of 140 word limit of tweets, identifying the polarity of statement will be a challenge.

Architecture/Components

The system consists of five phases.

- The first phase is data collection using streaming API.
- Second phase is Geocoding and Reverse Geocoding for identifying the state information of each tweet.
- Third phase is for identifying the popularity of car brands using the total number of tweets for a particular brand.
- Fourth phase is performing sentiment analysis on the tweets and identify the positive and negative polarity of tweets.
- Fifth phase is the visualization of data.

Change in Initial plan

Since the data collection using streaming API has limitations for analyzing the trends in previous data, along with the streaming API we will be using Rest API for collecting data in a time period. Data collected using Rest API will be used to identify the trends in car brands and will provide a visualization on the data. Top five brands in particular state in a particular year will be visualized. And Top five states in which a particular brand in a time period also will be visualized.

Evaluation Plan.

Four visualizations will be done on the twitter data for the car trends. Two for the current data using streaming API and two for the history data using Rest API. A clear distinction in trends in each state over a period of time can be considered as a successful implementation of the project MTV.

Issues.

Tweet Location: The major issue faced during the initial work is the data field "coordinates" which will provide the geo location of the tweet. It is missing in most of the tweets.

Issue is resolved by selecting another field Location which is the data user provided manually while creating the profile. Genuineness of this data is doubtful, because it is not necessary that he changes his location even if he moved to another location.

Geocoding and Reverse Geocoding: Another issue faced during the work is the query limit for geocoding and reverse geocoding. It is 2500/day or 5/sec.

The limit 5/sec is resolved by limiting the request by using sleep feature, but not resolved the limit of 2500/day. It can be resolved by faking the IP address. This will be achieved in the next phase.

Data encoding: Data Encoding for the twitter response varies with tweets. Sometimes the tweet has multi-byte characters and which might be difficult to perform sentiment analysis on data.

This issue is resolved to a limit using utf-8 encoding but not able to resolve it completely.

Initial Implementation

- Completed the initial phase for collecting the data using twitter streaming API.
- Completed the second phase for performing geocoding and reverse geocoding on the data to collect state information.
- Issue with the query limit is yet to resolve.
- Designed the web page layout for the final output.

Language: python 2.7
Libraries: Tweepy, pygeocoder
Tools: Stanford Named Entity Recognizer, D3.js

Algorithm

Popularity count: Popularity measure will be calculated by the count of car entities using the tool Stanford Named entity Recognizer and the count for each brand in each state will be calculated.

Sentiment analysis: Sentiment analysis will be performed in three phases.

Phase1: All emoticons will be replaced with the meaning using an emoticon dictionary which is having all emoticons and its meaning by using Wikipedia.

Phase2: All words will be assigned a pleasant score by using the DAL dictionary, which has score for each word ranging from 1 to 3.

Phase3: Using Naïve Bayes classification, sentiment analysis will be performed on all the tweets.

Trends: Trends will be calculated by using the popularity count together with positive and negative polarity. Positive polarity will be added to the polarity count and negative polarity will be deducted from the polarity count.

Tasks to be completed:

- Calculate the popularity count.
- Performing sentiment analysis on data.
- Visualization of data.

The project will be completed in the above mentioned three phases. The final output will be having four visualization**.** Two for the current data using streaming API and two for the history data using Rest API. A clear distinction in trends in each state over a period of time can be considered as a successful implementation of the project MTV.

The main expected challenge is the amount of tweets that will be having location information.