MOTOR TWEET VOGUE (MTV)
Project
CSE 6339
Spring 2015

- **Team information**
  Minumol Joseph.
  Jisa C Sebastine.

- **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.

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

- **What datasets you will use and how/where to acquire the datasets?**

  Current data will be accessed using twitter streaming API and historical data will be accessed using the search API.

Twitter Streaming API	: https://dev.twitter.com/streaming/overview/connecting
Twitter Search API	: https://gnip.com/sources/twitter/#

- What do you plan to deliver at the end of the semester? How would your demo and system look like?

  Our effort is to provide a visualization of popular trends in the field of motorcar models in each states of the country in different periods of time. Current trends will be shown as well as an option to select the time period will also be provided. The results of comparison of trends in each states will be presented.

  The results obtained will be useful to analyze the progress and regress of several brands. It will provide a gist of interests of indigenous people regarding their automobile craze. It is a window for automobile companies to see their popularity across regions.

- What are the challenges in this project? Why is it significant?

- Since streaming API is being used, it is not possible to predict the amount of data that will be used for the analysis .
- Named Entity recognition process is challenging because the same name is there for more than one type of entity(eg:Avenger, Rogue).
- Same name can be given in different ways(eg: Nick names)
- Sparsity of location Information in the data will be another challenge. Most of the data might not have location information along with tweet.
- Performing sentiment analysis will be another challenge.

- How do you plan to address the challenges? How would you design and implement the solution?
  - Size of streaming data - Gzip compression may reduce the bandwidth needed to process a stream to as small as 1/5th the size of an uncompressed stream. Requesting a gzipped stream might help to deal with the size of streaming data.
  - Entity recognition disambiguation - Use a dictionary of expected names and features of other words in that particular sentence also will be included for entity recognition.
  - Sentiment Analysis - Different algorithms will be used for the effective sentiment analysis. Better performance algorithm will be used for the successful project implementation. (Ref: Sentiment Analysis of Twitter Data, Department of CSE, Columbia University)

- How would you evaluate your project? How would you call it a success?

  If we are able to visualize the current motor car trend in each state and previous trends in that state, then the project is considered to be a success. If we are able to overcome the above mentioned challenges more than 50 %, then we take it as a success.

- How would you partition the tasks and coordinate among group members?

  Project will be completed in four phases as shown below. The person assigned to a particular task will be responsible for that task and the other person will support the same. Each task is divided into different phases based on its complexity.

  A tentative plan is provided below.

1) Streaming data collection and Reverse geocoding		Minu
2) Pattern Mining.	`	jisa
3) Sentiment analysis (2 ways)		
	Phase1	Minu
	Phase2	Jisa
4) Web development & Visualization		Jisa & Minu