CS 5590 Project Proposal

Team 1

Project Title:

Twitter Analysis of Streaming Data

Team Details:

Name	ClassID	Email
Acikgoz, Mehmet	1	ma96f@mail.umkc.edu
Attaluri, Lalith Chandra	4	la4kf@mail.umkc.edu
Karumanchi, Pranitha Saroj	7	pkt59@mail.umkc.edu
Wolfe, Jonathan Andrew	17	jawhf4@mail.umkc.edu

Overview

Perform an analysis of streaming twitter data. Data will be collected using hadoop and transferred to Hive using Sqoop for filtering. Data will be filtered to remove entries such as deleted tweets and refine tweets to remove retweets, replies and some languages. After the data has been filtered, it will be processed using Spark and GraphX to perform geotagging, bot detection, sentiment analysis and hashtag word relationships. We will also be doing visualization on the data. We divide into two parts they are:

MapReduce Framework:

One of the questions we had in our mind in this project is to find out the people tweeting more than others. In order to figure out the people who are using twitter, we decided to build a framework which can enable us to work with huge data set. Thus, we employed a java-based Mapreduce framework to perform the analysis required for finding the trends in Tweeters.

As we all know, a tweet object has the user object which contains information about the owner of the tweet. We decided to create a design which will help us find out the number of tweets by user within a time interval. For doing it, we used java-json library in the framework in addition to default cloudera hadoop distributed file system.

So, in order to find the top trends in tweeters in a given snapshot, we would need to:

1. Process all tweets and parse out tokens with "user.id_str"
2. Count all the 'user.id_str's.
3. Find out top n 'user.id_str's by sorting them

Sentiment Analysis of Tweets in Hive:

Here we are intended to perform Sentiment Analysis on tweets which we have extracted from Twitter using hive.

As the tweets extracted from twitter are in JSON format the biggest challenge we face here is to load the JSON data into Hive. To acheive this purpose we will utilise Hive JSON SerDe

Visualization

Tableau is a Data Visualisation tool that we are going to use for this project. It helps create interactive graphs and charts in the form of dashboards and worksheets to gain more insight. We will link Tableau to the data source to get started