Identifying Popular Events in a Region Using Extracted Twitter Data.

Introduction:
In the present day social networking data has gained huge prominence as it has got wide range of practical applications. One such is using it to make assumptions and predictions about user behavior. In our project we are planning to extract data (tweets) from “Twitter” social networking website and run analysis on the data to identify the events that are most discussed by its users. Once we have this analysis, the result can be used for advertising, marketing, etc.

Plan:
The plan mainly involves two phases:
1. Extracting the data from twitter.
   - This should be done over a period of time, so tweets published in the twitter should be extracted and stored daily for a period of 2 to 3 weeks. This will be our dump for performing the experiments.

2. Running Analysis on this data to identify the popular events in a region (country) over a period of time.
   - Analyze the extracted data using Hadoop and Map Reduce to get the statistics (popularity) of various events within a region.

Overview of Implementation:
The following steps illustrate how we are going to implement this project:

1. Extracting data from twitter.
   - For this project we are planning to extract only filtered data from twitter i.e. data that corresponds to a particular event, which will be the system input.
   - We are using “Twitter4j” java API for twitter to extract the data.
   - We will use a set of records as guidelines in order to address Name ambiguity (for details refer the Challenges Section)

2. Use the Geoname webservice to get the Country of the tweet.
   From the extracted tweet, we will be getting the longitude and latitude of tweet. We will use Geoname webservice to identify the Location of the Tweet.

3. Storing Tweet: Location in HDFS.
   - Once we have the extracted data from twitter we store the data in HDFS
Identifying popular events using extracted twitter data
Dileep Bodanki
Sagar Sugandhi

(Hadoop Distributed File System) to facilitate the analysis of the data. We store the tweets in text file in “Tweet” : “Location” format.

4. Running Map Reduce job to get “Country: No of Tweets “pattern.
   - We will write Map Reduce job to extract this information from the tweet.

Challenges:
Some identified challenges in our process are:

1. Name ambiguity in tweets.
   - When a word is used by a user in a tweet, it may imply something different than the word we are analyzing. For instance, some words like cricket have multiple meanings. Cricket is a sport as well as an insect.
   - So, for identifying the context of the tweet, we will also give a List of words that the tweet should contain along with “cricket”, to make it eligible for selection.
   - Currently, for the purpose of this project, we will be giving manually generated words.

2. Location identification.
   - Some tweets don’t return a location so this could be a problem if the number is large.

Demo and Evaluation:
Our system will work on the data collected from the twitter over a period of 2-3 weeks and will give as output the location and the number of tweets generated from it. So, we will be showing the final results and the demo of the individual processes like extraction and Map reduce Job.

   For Evaluation, we compare our analysis with market surveys (viz. TRP ratings etc) and can calculate how accurate our method is. We are planning to generate graphs from the data generated for a particular location and its weekly trends, and it would help us evaluate the system better, from the above surveys.

Future Plans:
In our current project we are analyzing the user traits for a single event but in the future we could extend this to multiple events. The system would extract unfiltered tweets and output the list of most popular events in each country using the extracted data. Also use of DBPedia knowledgebase to address the name ambiguity (refer Challenges section). If time permits, we would try to implement this in the term project.