Mine & Follow

The project titled 'Mine & Follow' will mine into the tweet data of millions of Twitter users & look out for 'relevant' Twitter users who should be followed. The idea is to carry a sentimental classification of tweet data and figure out their corresponding tweeter (Twitter user), so that they could be made aware of the subjects of their interest more extensively & simultaneously extend the reachability of a Twitter user who wishes to convey these new subjects of interest to some 'relevant' users.

Implementation Plan:

Our plan to implement this idea is as follows:

- Decide upon a specific industry, for ex: Automobile or Sports or Politics.
- Build a set most frequent/common keywords specific to that industry. For ex: Audi, BMW, etc., if Automobile is chosen. Cricket, Baseball, etc., if Sports is chosen.
- Access the Twitter dataset files (JSON format) and search for above keywords within the complete tweet data. This forms one job task, which will be handled by Hadoop framework in a map-reduce phase. The mapper phase will map each keyword with a corresponding tweeter name and emit. The reducer will collate over a list of values of tweeter names corresponding to that keyword.
  In the second Map-Reduce phase (chained), the Tweeter name in the Twitter data set will be searched for each of the Tweeter names found in step 1. We now retrieve the list of tweets for each tweeter and apply the same keyword as in step 1, along with other set of sentiment-predicting keywords such as- 'Good','Beautiful','Love','Worst','Bad','Never', etc. These sentiment-predicting keywords help filtering those tweeters, who are interested about the subject & would possibly like to receive more information on the same. So the classification is essentially a binary classification, i.e. 'Positive' & 'Negative'. The mapper in this second step will produce only positively classified Tweeter name as key and his e-mail ID/or any available contact info as value. We should not require any reducer in this phase, as the information within mapper will suffice our interest.
- A dummy Twitter user will follow the list of Tweeter names populated above. We will create this dummy user and process to make it follow the populated list of users will be achieved using Twitter API.
  The user contact info (populated as value field above) will be used to directly contact (preferably e-mail) the user.

The dataset to be used for this project is the Twitter dataset, which has been made publicly available in compressed JSON format.

Group Member names:

1. Ankit Malhotra (axm125230@utdallas.edu)
2. Kaushik Sirineni (kxs124530@utdallas.edu)
3. Venkata Mahesh (vxc121530@utdallas.edu)
4. Karthik Sanakayala (sxs129531@utdallas.edu)