**********README**********

Please run run.sh as:

sh run.sh

parseTweets.py : Parse the crawled tweets(json) and generate 'tweets.xlsx'

preprocess.py: Extract features(tweets text and http link) from tweets.
Use TF-IDF to convert text to vector

classification.py: Use logistic regression to classify tweets into Sport_Related and Non_Sport_Related.

dbscan.py: Use TruncatedSVD to reduce dimensions on features(from 4000 d to 200 d), then cluster Sport_Related tweets with DBSCAN.

gmm.py: Use TruncatedSVD to reduce dimensions on features(from 4000 d to 200 d), then cluster Sport_Related tweets with GMM

parseCountData.py: Parse total counts of each tweet, including quote_count, reply_count, retweet_count and favorite_count. This feature is used in Ranking Event step.

clusterDescrip_and_ranking.py: Each cluster represents an event. Generate summary of each cluster by selecting the highest-counts-tweet as the event summary of whole cluster. Then rank clusters according to each cluster's total counts.

The following two files are not included in run.sh.

crawlTweets.py: Use sport-related seeds to crawl tweets.

classificationEvaluation.py: Use cross-validation to evaluate the classification.