I collect data from twitter using the streaming API. The data set contains 38,040 records which includes 56744 related users (either the one posted tweets or the users who are mentioned in others’ tweets) and 6415 distinct hash tags.

The data not only help us to understand some basic statistics of current twitter, tweets and twitter users, but it can also be compared with data already available. By comparing current data and the sample data we already had, we can explore whether there are constant pattern for certain variable and what has changed most.

Although the API dataset is much smaller in size compared with sample dataset. As long as we assume the randomization of choosing data, we should expect the two datasets to be comparable. The results mainly confirm our hypothesis. The average numbers of tweets per user with multiple tweets are similar (2.18 vs. 2.32) in two datasets. Also, the max number of hashtags in tweets, the percent of overall dataset that has hashtags, the max number of user mentions in tweets, the percent of dataset that has user mentions are all similar in two datasets. However, we notice that the average number of hash tags and user mentions are quite different. For both variables, the API dataset has a much higher average number than the sample dataset (1.90 compared to 0.30; 1.33 compared to 0.80). While the percent of overall dataset that has hash tags or user mentions don’t changed in the two datasets, it seems that those tweets that have hash tags or user mentions are more likely to have more hash tags or user mentions and those tweets that don't are not influenced.