## Qualitative Analysis

Selecting top 1000 features and 100 after LSA

- 1. Perform K-means clustering with K = 5 for all clusters.
  - a. Found uniformly distributed results for Fs1. Fs2, 3 and 4 were skewed towards a couple of clusters.

Shubhangi: shtandon@ucsc.edu Sharath: sturuvek@ucsc.edu

- b. To fix this, we increased the number of clusters to 7 and 9. We observed that the cluster centers were still skewed towards the first two.
- 2. With 10000 features and 1000 after LSA
  - a. We still observed that one cluster was slightly skewed, and noticed that the tweets associated with this cluster were indeed similar and not skewed.
- 3. Goodness of clustering 10000 features, 1000 after LSA
  - a. With the feature set 1, we notice that tweets with similar wordings are clustered together. For examples, there is a cluster where all the comments on Hillary are negative.
  - b. There is not much difference in performance on feature set 1 in comparison to feature set 2. The number of points assigned to each cluster are comparable.
  - c. The feature sets 3 and 4 perform better when compared to the previous two. Clusters are not only based on similar wordings like it was previously.

Even with the four feature sets, there is considerable room for improvement which can be brought out by more discriminative feature extraction algorithms.

|             | tweet_id  | retweets  | cluster_fs1 | cluster_fs2 | cluster_fs3 | cluster_fs4 |
|-------------|-----------|-----------|-------------|-------------|-------------|-------------|
| tweet_id    | 1.000000  | 0.009350  | -0.102135   | 0.229265    | -0.163587   | -0.007626   |
| retweets    | 0.009350  | 1.000000  | 0.007296    | 0.010324    | -0.017222   | 0.011904    |
| cluster_fs1 | -0.102135 | 0.007296  | 1.000000    | -0.055285   | 0.259441    | -0.254475   |
| cluster_fs2 | 0.229265  | 0.010324  | -0.055285   | 1.000000    | -0.275981   | -0.011570   |
| cluster_fs3 | -0.163587 | -0.017222 | 0.259441    | -0.275981   | 1.000000    | -0.416079   |
| cluster_fs4 | -0.007626 | 0.011904  | -0.254475   | -0.011570   | -0.416079   | 1.000000    |

Table 1: Clustering correlation among four feature sets.

We noted that a set of tweets which have essentially the same text are grouped in the same cluster for each feature set, which is expected behavior. (See table 2 for a sample screenshot of data)

Shubhangi: shtandon@ucsc.edu Sharath: sturuvek@ucsc.edu

|       | tweet_id | text   | retweets | cluster_fs1 | cluster_fs2 | cluster_fs3 | cluster_fs4 |
|-------|----------|--|----------|-------------|-------------|-------------|-------------|
| 23594 | 1634013  | KhakieND #HillaryClinton fired for lies, unethical behavior        |          | 5           | 4           | 4           | 0           |
| 23506 | 1632357  | ForeverThenNow #HillaryClinton fired for lies, unethical behavior  |          | 5           | 4           | 4           | 0           |
| 45288 | 1632150  | KeithRPrior #HillaryClinton fired for lies, unethical behavior     |          | 5           | 4           | 4           | 0           |
| 5694  | 1643954  | ChadwickLois #HillaryClinton fired for lies, unethical behavior    |          | 5           | 4           | 4           | 0           |
| 35091 | 1682398  | Steven31015146 #HillaryClinton fired for lies, unethical behavior  |          | 5           | 4           | 4           | 0           |
| 18077 | 1645534  | drisnya #HillaryClinton fired for lies, unethical behavior         |          | 5           | 4           | 4           | 0           |
| 30610 | 1642329  | Dan2582Ortiz #HillaryClinton fired for lies, unethical behavior    |          | 5           | 4           | 4           | 0           |
| 13607 | 1636986  | pacholiejb #HillaryClinton fired for lies, unethical behavior      |          | 5           | 4           | 4           | 0           |
| 5742  | 1683072  | ReichelEric #HillaryClinton fired for lies, unethical behavior     |          | 5           | 4           | 4           | 0           |
| 30624 | 1637274  | MsRuthedelaRosa #HillaryClinton fired for lies, unethical behavior |          | 5           | 4           | 4           | 0           |
| 45241 | 1634437  | SallvJupiterRA #HillarvClinton fired for lies. unethical behavior  | 0        | 5           | 4           | 4           | 0           |

Table 2: Showing a set of tweets and the clusters assigned to them, w.r.t to the feature sets chosen.