Qualitative Analysis

Yanan Xie, Ziqiang Wang

We tried different numbers of cluster k from 5 to k.

In the case of k=5, more than 95% of tweets grouped into one cluster with the feature set 4(LSA dimension reduced features). Some clusters have only 1~3 tweets. We tried to run KMans and LSA several times with different numbers of latent semantics, however, this result kept appearing. After looking to the specific vectors generated by LSA, we found that the ranges of some positions of the vector are extremely small. So we switched the initialization algorithm of KMeans to random which provides us a larger chance to have relatively larger clusters rather than one super large cluster.

As the k increase, we found more tweets that are separating from the major group clustered with smaller k. So we choose the largest k in the given range which is 10 as our final result.

We take Clinton dataset as an example to illustrate some common features of some certain groups. Tweets that shares article Hillary Clinton fired for lies, unethical behavior (https://t.co/psXVPVrlTg) are grouped into one cluster(labeled 7,3,5,8 by four feature sets respectively). However, we do find some tweets that are labeled 8 by the feature set 4 and labeled differently by other feature sets. For example, tweet Trump, you're fired. #DumpTrump #ImWithHer #ClintonKaine2016 https://t.co/QsWYQHfZyG(2503291) is labeled 4,5,3,8. We can see it is grouped into the same group as previous one by the feature set 4 for the reason that it also talk about firing someone. We look into other tweets that are also labeled 4 by the feature set 1 and find that those tweets are just related by some common tokens like you’re. This also applies to groups labeled 5,3 by the feature set 2 and 3 which provides us a clue that the feature set 4 may be better than other feature sets. Another positive example for the feature set 4 is that almost all tweets with label 1 are talking about Clinton’s supporters.

However, most small groups are just tweets sharing the same articles. This indicates that tweets of small size fail to provide enough information to detect their topic relation. In conclusion, although the whole pipeline is increasing the quality of clustering, the final result is bad at finding topics.