# MACHINE LEARNING ASSIGNMENT 4

## TF-IDF REMOVE COMMON WORDS

In assignment work, irrelevant words are removed by filter out the stop words.

In TF-IDF case, since 'the' is a common word in text: although text frequency makes its weight increase, inverse document frequency makes total weight of common word in tf-idf become small since the common word also appears in many different document (titles).

$$Tfidf = tfxidf = \frac{word\ freq\ in\ doc}{total\ words\ in\ doc} \times \log \frac{number\ of\ documents\ in\ copus}{number\ of\ document\ contain\ word}$$

## VISUALIZE DATA

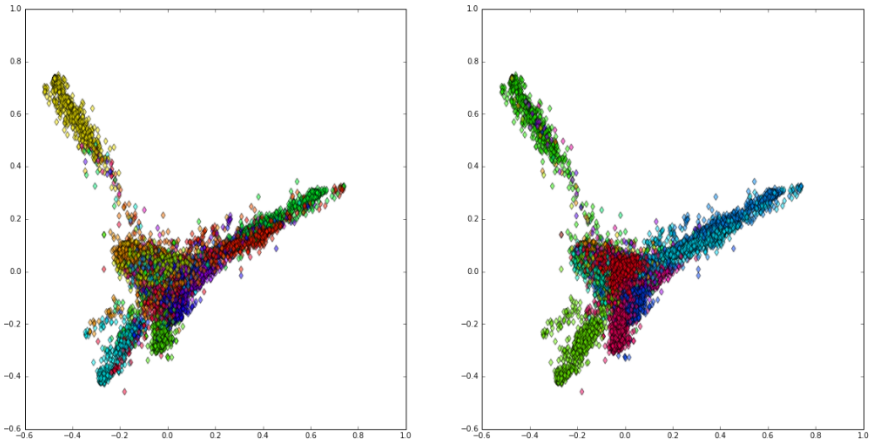

Figure 1 Cluster with 50 classes and true label with 120 components for SVD

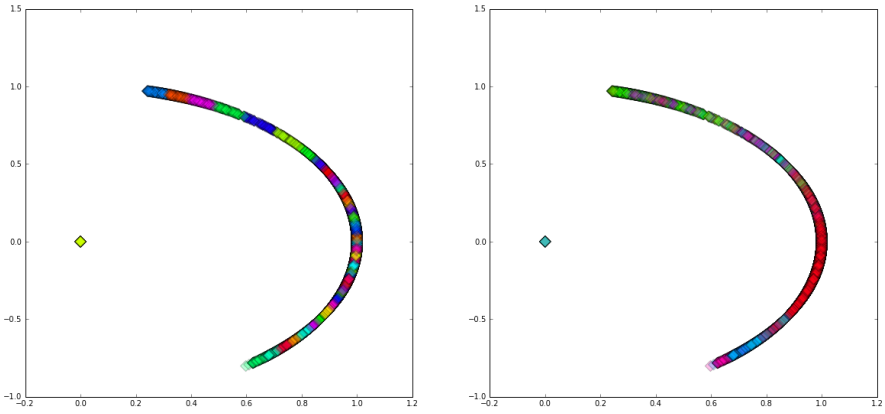

Figure 2 Cluster with 40 classes and true label with 2 components for SVD

In best score in this task is 50 clusters with 22 components for SVD, the result visualized data is mess in 2d representation form and less messy in 3d form. To conclude, clustering may work well in those high dimension environment but it is difficult to present in low dimension form.

## COMPARE DIFFERENT FEATURE EXTRATION METHODS

| Method | Score |
|---|---|
| BOW + Kmean clustering | 0.23327 |
| Word2Vector with sampling | 0. |
| TD-IDF + LSA + Kmean clustering | 0.51431 |
| | 0.60645 |
| TD-IDF + LSA (with more dimension) + Kmean clustering | 0.62434 – 0.80216(Post-Deadline) |
| Word2Vector | - (too long for process similarity matrix) |
| W2V (Sampling) + LSA | 0.39236 |

Since create full similarity matrix for topic cross similarity in word2vec method, a few topic is selected as sample to compute the distance with other topic as feature. Then clustering after dimension reduction.

## DIFFERENT CLUSTER NUMERS