Proposal

1. **Purpose :**

Cluster documents (news) from news website into different category using K-means mining technique.

1. **Dataset Description:**

This dataset is text collection from <http://www.dailykos.com/#.> This is a news website. After tokenization and removal of stopwords, the vocabulary of unique words was truncated by only keeping words that occurred more than ten times. We have 3430 documents collected from website, 6906 words in vocabulary, and 467714 words in the total collections. For this text collection, there are two files. Docword.\*.txt(the bag of words file in sparse format) with three attributes: Docid, Wordid, Count, 353160 rows in total. And vocab.\*.txt (the vocab file) with only one attribute: Word corresponding with Wordid, 6906 words in total. These data sets have no class labels, and for copyright reasons no filenames or other document-level metadata.

1. **Initial data preprocessing**
2. Consider the number of words is large, we need select meaningful words for clustering the documents. First of all, for each word, we calculate the number of occurrence of the word with plusing the counts the word appeared in each documents. A plot showed below can help us set a threshold for occurrence of words.
3. Set threshold as ‘Occurrence times’ equals to 450, then we left 106 words for clustering. And change attribute ‘wordid’ with corresponding words.
4. After we add three attributes header and load into weak, we need denormalized ‘wordid’. Then we got 107 attributes(including 106 words and ‘docid’).

4. **Description of the particular problem**

1. We have 3430 documents and 6906 words which is a quite large number to analysis, and it hard to load into weka. So, We need to select words with high frequency occurrence. As for this dataset, it is a problem to select threshold of occurrence times of words. 2. Besides, since this dataset has no class, for clustering, we can’t do classes to cluster evaluation. Only we can got from clustering dataset is number of clusters and documents(docid) each cluster have. 3. Additional, It is a problem to select the number of clusters.

**5. Ways to tackle problem mentioned above**

1. Draw a plot ‘times of occurrence’ vs ‘selected number of words’. We can see the curve and select a threshold. 2. Since it is a news website, we may cluster it into politics, environment, economy, etc. 3. For number of cluster, using silhouette plot, through select different clusters, we can get different silhouette value for each cluster. We choose the number of clusters which have silhouette values most close to 1 relatively.