Tweets Clustering

Twitter provides a service for posting short messages. In practice, many of the tweets are very similar to each other and can be clustered together. By clustering similar tweets together, we can generate a more concise and organized representation of the raw tweets, which will be very useful for many Twitter-based applications (e.g., truth discovery, trend analysis, search ranking, etc.) Here, the tweets are clustered using Jaccard distance metric and K-means clustering algorithm.

Jaccard Distance (Explanation)

The Jaccard distance, which measures dissimilarity between two sample sets (A and B).

It is defined as the difference of the sizes of the union and the intersection of two sets divided by the size of the union of the sets. Dist(A, B) = 1 - |A intersection B|/|A union B|

For example, consider the following tweets:

Tweet A: the long march Tweet B: ides of march |A intersection B | = 1 and |A union B | = 5, therefore the distance is 1 – (1/5) Jaccard Distance Dist(A, B) between tweet A and B has the following properties: 1. It is small if tweet A and B are similar.

2. It is large if they are not similar.

3. It is 0 if they are the same.

4. It is 1 if they are completely different (i.e., no overlapping words).

Dataset: https://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter

Main steps:

Tweets Preprocessing:

- tweet ids and timestamps are removed.
- words that starts with the symbol '@', e.g., @AnnaMedaris, are removed.
- hashtag symbols are removed, e.g., #depression is converted to depression.
- any URL are removed.
- every word is converted to lowercase.

K-Means Clustering Algorithm

K-means clustering algorithm is implemented from scratch, without using any machine learning libraries.

1) The code uses "bbchealth.txt" by default for the tweets data. - A user can change the url path to another data file as desired from the given files.

2) The code uses, "3 clusters" by default and performs "5 experiments" one after another.

3) user can change the default value of initial clusters (k) and number of experiments to be performed. 4) The program returns the value of SSE (sum of squared error) and size of each cluster after every experiment (plotted).