CS 6375 Project-3

Report on KNN clustering on Twitter data

Submitted by - Yash Vijaynarayan Gupta

Steps of the exercise:

(1) We are going to use the following dataset for this exercise:

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

Follow the "Data Folder" link and unzip the given file. You will file a folder containing tweetsthat contain links to various news sources e.g. the file "usnewshealth.txt" contains tweets that refer to articles published in US News. **You have to choose one such file and proceed. I have used "cnnhealth.txt".**

(2) Perform the following pre-processing steps (may vary slightly for different files in folder depending on diversity of data inflections):

- Remove the tweet id and timestamp
- Remove any word that starts with the symbol @ e.g. @AnnaMedaris
- Remove any hashtag symbols e.g. convert #depression to depression
- Remove any URL
- Convert every word to lowercase

(3) Perform K-means clustering on the resulting tweets using at least 5 different values of K andreport your results in the format below
Note that the sum of squared error is defined as:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

where K is the number of clusters and m_i is the centroid of the i^{th} cluster.

Results:

K-Values	Square Sum Error Values	Cluster Sizes
3	540.1677557435883	Cluster 0 : 2271 tweets Cluster 1 : 1709 tweets Cluster 2 : 81 tweets
4	517.1981082077518	Cluster 0 : 1882 tweets Cluster 1 : 806 tweets Cluster 2 : 758 tweets Cluster 3 : 615 tweets
5	526.9556327921812	Cluster 0 : 1199 tweets Cluster 1 : 25 tweets Cluster 2 : 933 tweets Cluster 3 : 1737 tweets Cluster 4 : 167 tweets
6	475.3478101779239	Cluster 0 : 373 tweets Cluster 1 : 1125 tweets Cluster 2 : 354 tweets Cluster 3 : 776 tweets Cluster 4 : 770 tweets Cluster 5 : 663 tweets
7	490.0386572055885	Cluster 0 : 424 tweets Cluster 1 : 256 tweets Cluster 2 : 471 tweets Cluster 3 : 1743 tweets Cluster 4 : 168 tweets Cluster 5 : 806 tweets Cluster 6 : 193 tweets

per page

We got the result on 5 different k values ranging from (3-7)

The number of tweets that were grouped together in each scenario can be seen in the table above. Clusters numbered from 0 onwards mentioned with their respective size of tweets.