CS 6375 - PROJECT 3
KNN Clustering on Twitter dataset

Tweets Clustering using k-means

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.)

In this assignment, you will learn how to cluster tweets by utilizing Jaccard Distance metric and K-means clustering algorithm.

Objectives:

□ Compute the similarity between tweets using the Jaccard Distance metric.
□ Cluster tweets using the K-means clustering algorithm.

Jaccard Distance:
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.

$$Dist(A,B) = 1 - \frac{|A \bigcap B|}{|A \bigcup B|} = \frac{|A \bigcup B| - |A \bigcap B|}{|A \bigcup B|}$$

For example, consider the following tweets:

Tweet A: the long march

Tweet B: ides of march

|A ∩ B | = 1 and |A U B | = 5, therefore the distance is 1 – (1/5)

In this assignment, a tweet can be considered as an unordered set of words such as {a,b,c}. By "unordered", we mean that {a,b,c}={b,a,c}={a,c,b}=...

Jaccard Distance Dist(A, B) between tweet A and B has the following properties:

□ It is small if tweet A and B are similar.
□ It is large if they are not similar.
□ It is 0 if they are the same.
□ It is 1 if they are completely different (i.e., no overlapping words).

Here is the reference for more details about Jaccard Distance:
http://en.wikipedia.org/wiki/Jaccard_index

Hint: Note that since the tweets do not have numerical coordinates as in Euclidean space, you might want to think of a sensible way to compute the "centroid" of a tweet cluster. *This could be the tweet having minimum distance to all of the other tweets in a cluster.*

Exercise:

Implement the tweet clustering function using the Jaccard Distance metric and K-means clustering algorithm to cluster redundant/repeated tweets into the same cluster. **Remember that you have to write your own code for K-means clustering**. It is acceptable to use external libraries for *data loading and pre-processing only*. Python is the preferred language for this assignment. If you want to use any other language, clearly specify how to compile and run your code in the README file.