Grading Criterion of Assignment 3

Total Score: 100 points

Extra Credits: 20 points

Part 1: Tweets Clustering (100 points):

Task:

Implement the tweet clustering function using the Jaccard Distance metric and K-means clustering algorithm introduced above to cluster redundant/repeated tweets into the same cluster. You are expected to do the K-means implementation by yourself, so please do **not** use any external library that has K-means implementation in your code.

What to Turn In:

(1) A result file that contains the clustering results. Each line represents a cluster. It is in the form of *cluster_id: a list of tweet IDs that belongs to this cluster*

(2) The source code to finish this task.

Grade Criterion:

1. Whether submitted codes can run successfully.
   - ■ Successfully: **30 points**
   - ■ Unsuccessfully:
     --Jaccard Distance Computation is Correct: 10 points
     --K-means Implementation is Correct: 10 points

2. Whether result file contain the correct cluster result.
   - ■ Correct Results: **70 scores**
   - ■ Partially correct results: The scores depend on how many correct clusters your found as well as the correctness of your code.

Part 2: Initial Seeds Selection (20 points):

Task:

Design and implement an *efficient* algorithm to find the k initial centroids so that the K-means algorithm you implemented can generate good clustering results (similar as the results you obtained using the 25 seeds we provided to you).

What to Turn In:
(1) A result file that contains the clustering results. Each line represents a cluster. It is in the form of *cluster_id: a list of tweet_id that belongs to this cluster.*
(2) The source code to finish this task.
(3) A README file that briefly explains the main idea and implementation of your algorithm to find the initial seeds.

Grade Criterion:

1. Whether submitted codes can run successfully.
   - ■ Successfully: **20 points**
   - ■ Unsuccessfully:
     -- Jaccard Distance Computation is Correct: 10 points
     --The Main Idea of Selection Algorithm is Correct: 10 points