Data Mining (CSE 5334) – Project Document 2

Mahesh Hingane
Student ID: 1001017122

Problem Scenario – The task is to cluster the questions from the site stackoverflow.com given the title and the text of the questions.

Project Specifications – The project is developed in *Python*. It takes the file (containing dataset to be clustered) and the number of clusters to be created as command line arguments. The result is generated into a new text file, with question numbers and the cluster id of the questions.

The project is based on *K-means* clustering method. The code first extracts all unique words from the data. Each of these words can be seen as a coordinate of the question vector. Then, the tf-idf value for each coordinate is calculated. These tf-idf values represent the vector form of the questions. These question vectors are then normalized, so that they can be compared by their distances. For the first iteration, *k* (number of clusters to be created) random questions are selected as centroids. Distance of each normalized question vector from each of these centroids is calculated, and the question vectors are placed in the closest centroid's cluster. For the next iteration, new centroids are calculated by taking average of coordinates of question vectors placed in each cluster. All the question vectors are then again checked for similarity with the new centroids and they are rearranged in the modified clustering scheme. The process of finding new centroids and reclustering the question vectors is repeated until the new centroids do not move any more.

Algorithm implemented:

1. Iterate through the input file to identify all unique words (coordinates).
2. For each coordinate, calculate the number of questions having that word. This will give the idf value for that word as -
   $$idf_{word1} = N \,/\, (\text{count of questions with word1 in them})$$
   where N is the total number of questions.
3. For each coordinate word, count the number of occurrences in each question containing that word. This will give the value of tf for each word in each question as-
   $$tf_{word1\text{-}que1} = 5*(\text{count of occurrences of word1 in title of question1}) + \text{count of occurrences of word1 in text of question1}$$
   Note: To have higher weight to words appearing in the title, each occurrence of word in title is weighted 5 times more than the occurrence in text.
4. Once the tf-idf values are available, normalize the question vectors by dividing each tf-idf value by the vector length.
5. Select *k* (number of clusters to be created) random questions.
6. Calculate the distance of each question from each of the k centroids. Put each question in its closest cluster.
7. Within each cluster, find a new virtual centroid with coordinate values as average of each coordinate of each question vector in that centroid.

8. If these new centroids are not exactly same as the old centroids, repeat the process from step 6.
9. Write the question ids and their respective centroid ids to the output file.

Comments -

1. In the project, only one iteration of the algorithm is implemented.
2. As of now, the project is executed with a smaller input file with 250 questions in it. The input file is submitted with the code. The result files for k=2, k=3, k=4 and k=5 are also submitted.