CS 6375 Assignment 3: K-Means clustering using Jaccard distance

Names of students in your group:

1. Mounika B (MXB210007)
2. Saketh Dasavathini(SXD190016)

Number of free late days used: 0

Theoretical part(Part 1): A separate pdf is attached consisting of steps of derivation for each of the problems given.

Programming part(Part 2):

Dataset used: "bbchealth.txt"- (https://raw.githubusercontent.com/bmounikareddy98/Machine-learning-assignments/main/Assignment3/bbchealth.txt)

Note: For the purpose of code execution we have already hosted the dataset on github public repository, hence not required to download it.

Description:

The dataset used is about tweets related to bbc_health. It is a text file consisting of wide range of tweets about bbc health. We have applied k_means clustering on the dataset using Jaccard distance as a distance metric and clustered the tweets into different groups. Below explains in detail about each parts and the steps involved.

Artificial neural network

1. Data preprocessing is performed
   Step 1: The \n at the end of each tweet is removed.
   Step 2: The tweet-id and timestamp are removed.
   Step 3: The words that start with "@" are removed.
   Step 4: The hashtags are removed
   Step 5: The URLs are removed
   Step 6: The words in the tweet are converted to lowercase.

   This ends the data pre-processing phase.

2. Performing K-Means clustering

A function to perform k-means is written. Initially random centroids are assigned. We have checked for the converge of k-means and the iterations are run accordingly. Functions to calculate Jaccard distance between two points(tweets) is written. Distance between the tweets and centroid is calculated and accordingly the points are assigned to each cluster. Later a function to update the centroids is written. The sum of squared errors is calculated for different values of k and number of tweets for different values of k is also outputted. **The SSE is minimum when the value of k is 6.**

Value of K	SSE	Size of each cluster
3	3401.05777	Cluster_1: 1301 tweets Cluster_2: 1029 tweets Cluster_3: 1599 tweets
4	3348.13454	Cluster_1: 605 tweets Cluster_2: 1470 tweets Cluster_3: 835 tweets Cluster_4: 1019 tweets
5	3343.73812	Cluster_1: 1239 tweets Cluster_2: 564 tweets Cluster_3: 532 tweets Cluster_4: 682 tweets Cluster_5: 912 tweets
6	3302.69309	Cluster_1: 538 tweets Cluster_2: 399 tweets Cluster_3: 1173 tweets Cluster_4: 755 tweets Cluster_5: 617 tweets Cluster_5: 447 tweets
7	3355.06509	Cluster_1: 357 tweets Cluster_2: 1225 tweets Cluster_3: 518 tweets Cluster_4: 557 tweets Cluster_5: 388 tweets Cluster_5: 437 tweets Cluster_6: 447 tweets