Tweets clustering project

Project Overview:

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.) Here, the tweets are clustered using Jaccard distance metric and K-means clustering algorithm

Standards:

- basics of AI
- Python programming language
- Algorithms of Machin learning & deep learning
- Methods of Machin learning & deep learning like (plotted).
- Dataset

Objectives:

- The code uses "bbchealth.txt" by default for the tweets data. - A user can change the URL path to another data file as desired from the given files.
- The code uses, "3 clusters" by default and performs "5 experiments" one after another
- user can change the default value of initial clusters (k) and number of experiments to be performed.
- The program returns the value of SSE (sum of squared error) and size of each cluster after every experiment (plotted)

Requirements/Task(s):

Task 1: using clean data
Task 2: using Jaccard Distance
Task 3: using k-Mean Clustering Algorithm

Our notes \ Research :

Use the clean code and link all the dataset files , link them with the Jaccard distance using intersection and union , and Learn how to build K-mean algo from scratch and link them with the files and show the plotted graph.

Outline the steps\plan of our project :

1. Using clean dataset
2. Build k-mean from scratch
3. Using Jaccard distance algorithm
4. Using plots and graph

Summarize what we learned :

How to build function from scratch , how to use Jaccard algorithm , how to clean exited dataset file txt , and using pandas ,re, string, math, and metaplot library