DSCI-560 Assignment No. 4 - Part 2
Instructor: Young Cho, Ph.D.

1) Forum Analysis & Clustering Algorithms

This assignment is the second part of a 2-part assignment that focuses on providing you with web scraping, data preprocessing, semi-automated topic selection, clustering algorithms, and real-time data processing. You will have a better understanding of how to collect and organize data from online forums and create a system for grouping related messages.

In this lab, you will work with your team to implement a clustering algorithm to group similar messages and display messages closest to the centroid of each cluster. Additionally, you will automate the data collection, processing and storage process to run at fixed intervals.

2) Message Content Abstraction

Convert messages into vector values of fixed dimensions that represent the meaning of the messages. This vector can be stored in the database along with the cleaned message.

One well-known method is Doc2Vec. Doc2Vec is a neural network-based approach that learns the distributed representation of documents. It is an unsupervised learning technique that maps each document to a fixed-length vector in a high-dimensional space.

https://www.geeksforgeeks.org/doc2vec-in-nlp/

There are other ways to accomplish similar functions and you are free to use a method that you deem sufficient for this lab. Please give your reason behind the method that you have chosen.

3) Clustering Messages

Develop a simplified algorithm for clustering the messages based on the keywords extracted from the text. If you are unfamiliar with the concept of clustering, please refer to the following tools.

- Scikit-Learn:Scikit-Learn Library Clustering
- NLTK Python:NLTK (Natural Language Toolkit)
- TextBlob: Documentation

Implement an algorithm to cluster the messages based on the vector extracted from the text. Identify keywords that are associated with all messages in each cluster. You may choose the libraries and toolkits or use other tools to cluster the vectors representing the meaning of the messages.

Use any visualization tool to display the final results, including N clusters of messages and their keywords. Verify that the contents of each cluster are similar by displaying and comparing message contents.

4) Automation

For this task, you will utilize the scripts from Part-1. Write an additional function that will call the web scraping, pre-processing and storage periodically to keep the database updated in real time.

Input an interval (in minutes) from the user as a command line argument of the form **"python filename.py 5"** and run the scripts and update the database after the given interval **(in the above example: 5 mins)** till the script gets an input string as **"quit"**. The script should provide proper messages such as fetching data, processing data, database updates, etc and appropriate error messages in case any of the operations fail.

When the script is not updating the database in the background, the command line prompt should take keywords or a message as input and find the cluster that matches closest to the input. The messages from a selected cluster should be displayed along with graphical representation.

5) **Team Discussions**

Your team is expected to meet in-person / virtually each day of the week and discuss the assignment progress & next steps. Document and compile minutes of all meetings in a separate file called **'meeting_notes_A4_P2_<team_name>.pdf'**

6) Submission

Make one submission per team. Each team must submit all the code files for the working solution, a readme document containing information for running the code in pdf format and a document that outlines the minutes of all team meetings in pdf format.

Provide a video per team which demonstrates the entire working solution and explains which algorithm was used and the rationale behind it. Also include details about the performance metrics used by your team. Please include the team name and the name of all three team members in the video.

There will be a 50% penalty for all late submissions.