DSCI-560-Lab5

Instruction

To run the pipeline, execute `main_pipeline.py` with a specified time interval (in minutes) as a command-line argument (e.g., `python main_pipeline.py 500` to update the database every 500 minutes).

The script follows these steps:

1. **Data Fetching**
   - Navigate to the `data_fetching` folder and Run `data_collection.py`, specifying the number of posts (default 5000) to fetch.
   - Don't forget to Initialize the `.env` file with Reddit client credentials and SQL database information and Create the necessary database by running the SQL commands in `sql.txt`.
2. **Data Processing**
   - Navigate to the `data_processing` folder and execute `data_processing.py`.
3. **Clustering**
   - Navigate to the `clustering` folder and Run `doc2vec.py` to process the data.
   - Don't forget to Execute the SQL commands in `sql_command.txt` to add two new columns to the `cleaned_posts` table.

Additionally, you can enter any message in the command line. The script will identify the closest matching cluster and display messages from that cluster, along with a graphical representation of word frequencies.

The results can be found in the meeting notes, which include a plot visualizing the K clusters (from K-means clustering) of messages and their associated keywords. Additionally, we provide sample titles and messages from posts within each cluster. By analyzing these messages, we observe that the clusters align with intuition—for instance, one cluster primarily consists of posts related to personal investment and advice-seeking, while another focuses on company trends and industry developments.