README: LAB 4 PART 2

Important Files and Folders

Ensure the following files are present in the root directory of the project:

1. `.env` - File containing environment variables such as API credentials for Reddit and database connections.

2. `requirements.txt` - List of dependencies to be installed for the project.

3. `database.py` - Handles the connection between the Python script and the MongoDB database.

4. `clustering.py` - Responsible for training the clustering model using Doc2Vec and KMeans, and visualizing the clusters.

5. `extract.py` - Fetches and cleans Reddit posts, extracts keywords using TF-IDF, and stores them in the database.

6. `automation.py` - Runs the periodic updates to scrape new posts from Reddit and cluster them in the background.

7. `query_cluster.py` - Allows users to search posts by keyword and find the most relevant clusters based on their input.

Setup

Follow these steps to set up and run the project:

1. Move into the project directory:

- cd reddit-scraper

2. Create and activate a virtual environment:

- virtualenv env
  source env/bin/activate # For Windows: "env\Scripts\activate"

3. Install the necessary libraries:

- pip install -r requirements.txt

4. Set up environment variables:

Create a `.env` file with the required environment variables such as:

- - REDDIT_CLIENT_ID
  - REDDIT_CLIENT_SECRET
  - REDDIT_USER_AGENT
  - MONGO_URI, MONGO_DB_NAME, MONGO_COLLECTION_NAME

This will allow the script to connect to Reddit and MongoDB.

Running the Script

Make sure all the important files are in place and the virtual environment is active.

To run the periodic updater script:

- Run the `automation.py` script:
  python automation.py

- Enter the subreddit name when prompted.
- Provide the time interval in minutes for periodic updates.

To run the query and clustering script:

- Run `query_cluster.py`:
  python query_cluster.py

  - Enter a keyword or phrase to search for relevant posts and clusters.

Execution Flow

The scripts follow the below steps:

1. Fetching Reddit Posts:

- The `extract.py` script fetches Reddit posts from a specified subreddit, cleans the text, and extracts keywords using TF-IDF.

2. Clustering:

- The posts are clustered using Doc2Vec to generate document vectors and KMeans to group them into clusters. The clustering results, including cluster labels, are stored in the MongoDB database.

3. Visualization:

- The clusters can be visualized using either PCA or t-SNE for dimensionality reduction, displaying the posts in a 2D scatter plot.

4. Querying the Clusters:

- Users can query posts by keywords, and the script will return posts from the most relevant cluster. If no direct match is found, the script finds the closest match using document similarity.

5. Automation:

- The `automation.py` script periodically fetches new posts from the subreddit, clusters them, and updates the database.

Technology Stack and Reasons

Python
Python was chosen for its extensive libraries in data scraping, machine learning, and natural language processing (NLP). Libraries like `gensim`, `scikit-learn`, and `pandas` simplify tasks like vectorizing text and clustering.

MongoDB
MongoDB is used as the database because it handles unstructured data efficiently, which is ideal for storing raw Reddit posts and their associated metadata. Its flexibility allows quick updates with new posts and their clusters.

Doc2Vec and Gensim
Doc2Vec (from Gensim) is used to create vector representations of posts. This model is well-suited for understanding document-level similarities, making it ideal for clustering posts based on content.

KMeans (from scikit-learn)
KMeans clustering is chosen for its simplicity and scalability. It's used to group similar posts into clusters, helping to discover thematic relationships between posts in a given subreddit.

t-SNE and PCA for Visualization
t-SNE and PCA are both used for reducing the high-dimensional vectors into 2D for visualization purposes. This allows the user to see a clear representation of the different clusters of posts.

Reddit API (PRAW)
The PRAW (Python Reddit API Wrapper) is used to fetch data from Reddit. PRAW simplifies interactions with Reddit's API, allowing easy access to subreddit posts.

TF-IDF for Keyword Extraction
The TF-IDF (Term Frequency-Inverse Document Frequency) method is used to extract meaningful keywords from each post. This helps in querying and understanding the posts more effectively.