ReadMe for Running Reddit Analysis Script

Introduction

The provided script is designed to scrape posts from a specific Reddit subreddit using the PRAW library. It further processes the posts, creates embeddings using the Doc2Vec algorithm, and then performs clustering using the KMeans algorithm. The results are then stored in a MySQL database.

Prerequisites

1. **Python**: Ensure you have Python installed on your system.
2. **Python Libraries**: The following libraries are required:
    - praw
    - pandas
    - datetime
    - time
    - re
    - nltk
    - mysql.connector
    - numpy
    - threading
    - warnings
    - sys
    - gensim
    - sklearn
    - wordcloud
    - matplotlib
    - tkinter
3. **MySQL**: Ensure you have MySQL installed and running on your system.
4. **Reddit API credentials**: You will need to create an application on Reddit to get `client_id`, `client_secret`, and `user_agent` to access the Reddit API.

Instructions

1. **Setting up PRAW**:
    - Create a Reddit account (if you don't have one).
    - Go to https://www.reddit.com/prefs/apps and scroll down to the "Developed Applications" section.
    - Click "Create App" or "Create Another App".

    - Fill out the required information and click "Create app" at the bottom when done.
    - After creating the app, you will get your `client_id` (right below the app name) and `client_secret`.
    - Set your `user_agent` to a unique and descriptive string.
2. **Script Configuration**:
    - Replace `client_id`, `client_secret`, and `user_agent` in the script with your respective values from the PRAW setup.
3. **MySQL Configuration**:
    - Make sure you have a MySQL server running.
    - In the `store_db` function, adjust the connection details (`host`, `user`, `password`) to match your MySQL server configuration.
4. **Running the Script**:
    - Navigate to the directory containing the script.
    - Run the script with the following command:

      ```
      python [script_name].py [period_in_minutes]
      ```

      Replace `[script_name]` with the name of the provided script and `[period_in_minutes]` with the desired period (in minutes) for updating the database.
    - The script will start fetching posts from the specified subreddit, processing them, and then saving them to the database.
    - For interactive analysis, the script will prompt the user to input keywords or messages. The provided input will then be clustered with existing posts to determine its relevancy.
5. **Stopping the Script**:
    - To stop updating the database and start the clustering, type 'quit' at the input prompt.
6. **Visualization**:
    - After providing input, the script will display various plots showing clusters and keyword distributions.

Notes

- Make sure to have the nltk data downloaded. The script includes `nltk.download()` functions for necessary data sets, but you can run them manually if needed.
- The subreddit to fetch from and the post amount are not provided in the visible part of the code. Make sure they are specified if you have the complete script.
- Always handle API credentials with care. Avoid hardcoding them directly in scripts that are shared or published.

Troubleshooting

1. **PRAW Authentication Error**: Double-check your Reddit API credentials.

2. **MySQL Connection Error**: Ensure your MySQL server is running and the connection details in the script are correct.
3. **Dependency Errors**: Ensure all required Python libraries are installed.
4. **Rate Limit Errors**: If you're making too many requests in a short period of time, Reddit may temporarily block your requests. Consider increasing the period or handling rate limits more gracefully.

For any other issues or further assistance, please refer to the respective library's documentation or seek online forums like StackOverflow.