README: Running the Reddit Post Scraper with Batch Processing

This script allows you to scrape new posts from a specified subreddit using the PRAW (Python Reddit API Wrapper). This script employs batch processing to efficiently fetch the desired number of posts. Once fetched, the posts are processed to extract relevant information and store it in a MySQL database.

Prerequisites:

1. **Python**: Ensure Python is installed on your system.
2. **Required Libraries**: The following Python libraries are required to run the script:
   - praw
   - pandas
   - datetime
   - re
   - nltk
   - mysql.connector
   - numpy

   You can install these using pip:

   ```
   pip install praw pandas nltk mysql-connector-python
   ```

3. **MySQL Database**: Ensure you have MySQL set up on your system. This script creates a new database and table to store the scraped data.

Instructions to Run:

1. **PRAW Account Setup**:
   - Visit the PRAW application preferences on Reddit and create a new application. Note down the `client_id`, `client_secret`, and set the `user_agent` (already set in the script).
   - Update the `client_id`, `client_secret`, and `user_agent` in the script with your details.
2. **MySQL Credentials**:
   - Update the MySQL connection details, particularly the `user` and `password`, with the appropriate credentials.
3. **Running the Script**:
   - Navigate to the directory containing the script using a terminal.
   - Run the script using the following command:

   ```
   python <script_name>.py
   ```

- Replace `<script_name>` with the name you've saved the script as.
- The script will fetch posts in batches, ensuring efficient processing especially for large numbers.

4. **Output**:
   - The script will display the scraped data after cleaning and after keywords extraction.
   - Data is also stored in a MySQL database named `Lab4_NAH` in a table named `reddit_posts`.
5. **Notes**:
   - This script scrapes from the 'tech' subreddit. To change this, modify the `subreddit` variable.
   - The script employs batch processing to ensure efficient fetching of a large number of posts. Each batch is limited to a maximum of 1000 posts.
   - Ensure you have enough API rate limits available for PRAW if you're trying to scrape a large number of posts.