**Lab 5: Web Scraping, Data Processing, and Clustering**

**Team Members:**

* Pratham Solanki
* Jaival Upadhyay
* Mayank Patil

**Team Name: Guardians of the Algorithm**

**1) Introduction**

This lab focuses on web scraping, data preprocessing, forum analysis, and clustering. The goal is to collect data from Reddit, clean and preprocess it, and apply clustering techniques to categorize discussions based on similarity. The final implementation stores and processes the data for meaningful insights.

**2) Data Collection & Storage (Implemented in Lab5-1.py)**

This module is responsible for:

* Scraping Reddit posts using the PRAW API.
* Filtering out irrelevant or promotional posts.
* Cleaning raw text data by removing HTML tags, special characters, and extra spaces.
* Extracting keywords from text using stopword filtering.
* Performing OCR (Optical Character Recognition) on images to extract embedded text.
* Storing data in a MySQL database, ensuring efficient handling of large data requests.
* Handling API limits using time-based pagination to fetch large amounts of posts within the allowed time frame.
* Exporting collected data into a CSV file for further analysis.

**3) Data Preprocessing (Implemented in Lab5-1.py)**

* Username masking: Usernames are anonymized to protect privacy.
* Text cleaning: HTML tags, special characters, and unnecessary whitespaces are removed.
* Keyword extraction: Stopwords are removed, and significant words are extracted as features.
* OCR processing: Image text extraction is performed using pytesseract.
* Data storage: Cleaned text, extracted keywords, and additional metadata are saved in a structured MySQL database.

**4) Forum Analysis & Clustering (Implemented in Lab5\_clustering.py)**

This module is responsible for:

* Loading preprocessed Reddit posts from the CSV file.
* Generating document embeddings using Doc2Vec, which transforms text into vector representations.
* Clustering messages using K-Means with an optimal number of clusters determined by the Silhouette Score.
* Identifying the top keywords for each cluster to provide insights into topic distributions.
* Visualizing clusters:
  + PCA (Principal Component Analysis) is used to reduce dimensions and plot the clusters.
  + A histogram is generated to show the frequency of different clusters.
* Classifying new keywords: Given a set of new keywords, the script predicts their corresponding cluster based on similarity.
* Cosine distance computation: Measures the similarity between new input keywords and existing clusters.

**5) Automation**

(Leave this section blank as requested)

**6) Execution Instructions**

**Prerequisites**

Ensure the following dependencies are installed before running the scripts:

pip install praw mysql-connector-python beautifulsoup4 pytesseract nltk pandas numpy matplotlib seaborn gensim scikit-learn

Additionally, ensure that nltk stopwords are downloaded using:

import nltk

nltk.download('stopwords')

**Running the Scripts**

**Step 1: Data Collection and Storage**

Run Lab5-1.py to scrape and store Reddit data.

python Lab5-1.py

**Step 2: Clustering Analysis**

Run Lab5\_clustering.py to perform clustering and visualization.

python Lab5\_clustering.py

**7) Conclusion**

This lab provides hands-on experience in web scraping, data preprocessing, and clustering algorithms. By structuring data from online forums and identifying clusters, we gain valuable insights into discussion trends and topic segmentation.

8) Github History