README

DSCI-560 Lab 5

Team Name: Trojan Trio
Team Members: Jinyao Yang (4266900395), Sen Pang (8598139533), Qianshu Peng (5063709968)

Github Link: YJY4266900395/USC-DSCI-560-lab5

1. Environment Requirements

Operating System

Ubuntu / Linux (recommended)

Python Version

Python 3.9 or above

2. Required Python Packages

Install required libraries:

pip install requests sentence-transformers scikit-learn numpy pandas matplotlib mysql-connector-python tqdm pytesseract

If OCR is not required, pytesseract can be omitted.

3. MySQL Setup

Ensure MySQL server is running.

Set the following environment variables:

export DB_HOST=localhost
export DB_PORT=3306
export DB_USER=your_username
export DB_PASS=your_password

export DB_NAME=dsci560_lab5

The ingestion script will automatically create the required table if it does not exist.

4. How to Run the System

The system consists of five stages.

Step 1: Scrape Reddit Data

Example command:

```
python3 fetch_reddit.py 5000 \
--subs tech cybersecurity technology artificial datascience computerscience \
--sorts new hot top rising \
--out_prefix posts_lab5 \
--checkpoint ck_scrape_5000.json \
--ocr \
--ocr_budget_images 5000 \
--ocr_max_images_per_post 3
```

This command:

- Collects 5000 posts
- Rotates across multiple subreddits and sort modes
- Enables OCR (optional)
- Saves checkpoint file for recovery

Output file:

posts_lab5_5000.jsonl

Step 2: Load Data into MySQL

python load_jsonl_to_mysql.py posts_lab5_5000.jsonl --table reddit_posts

This performs:

- Data cleaning
- Username pseudonymization
- Keyword extraction
- Upsert insertion into MySQL

Step 3: Run Embedding and Clustering

```
python embed_and_cluster.py \
--input posts_lab5_5000.json \
--k 8 \
--plot \
--write_db \
--table reddit_posts
```

This will:

- Generate semantic embeddings
- Perform K-Means clustering
- Save cluster centroids and labels
- Write cluster assignments back to MySQL
- Generate PCA visualization

Step 4: Interactive Query

```
python query_cluster.py "ransomware attack" \
--out_dir outputs \
--top_posts 5 \
--plot
```

This will:

- Encode the input query
- Match it to the most similar cluster
- Return representative posts
- Generate visualization

Step 5: Automation Mode

To run the system periodically:

```
python automation.py 5 \
--fetch_n 300 \
--k 8 \
--ocr \
--ocr_budget_images 300 \
```

--ocr_max_images_per_post 3 \
--table reddit_posts

This configuration:

- Runs every 5 minutes
- Fetches 300 new posts per cycle
- Performs ingestion, embedding, clustering
- Updates the database automatically

5. Output Files

The system generates the following outputs:

- centroids.npy
- labels.npy
- clusters_posts.csv
- clusters_summary.json
- PCA visualization files

Cluster assignments are stored in the MySQL table.