DSCI-560 Lab 5

https://github.com/herunkan/Data-Science-Professional-Practicum/tree/main/Lab5

Team Name: The Coach

Name	USC ID
Tianqi Qiu	1716906139
Shida Yan	5725964711
Herun Kan	7222919427

Initial Setup

This project was implemented in Python on a local macOS environment with MySQL used for persistent storage. The system was designed to perform web scraping, preprocessing, database storage, clustering, and automated periodic updates.

The primary libraries used in this project include:

- requests and beautifulsoup4 for web scraping
- pandas for data processing
- scikit-learn for TF-IDF embedding and K-Means clustering
- matplotlib for visualization
- mysql-connector-python for database interaction

Data Collection and Storage

Topic Selection

The initial target subreddit was r/Fitness. However, during preliminary data exploration, it was observed that r/Fitness contained a significant number of repetitive megathreads and lower topical consistency. As a result, the project was refocused on r/workout to obtain cleaner and more coherent content. All final analysis presented in this report is based on r/workout.

Scraping Method

Data was collected by crawling old.reddit.com using the BeautifulSoup library. Pagination was handled by repeatedly following the “next” links to retrieve additional posts, avoiding reliance on single-page limits. The scraping script accepts a command-line parameter specifying the number of posts to collect, allowing flexible experimentation with dataset size.

Database Storage

All collected posts were stored in a MySQL database named lab5_workout. The primary table, reddit_posts, stores post-level data, while a supporting table, cluster_keywords, stores representative keywords for each cluster.

An upsert mechanism was implemented to ensure that repeated runs update existing rows instead of inserting duplicate entries. This design supports automation and periodic updates without corrupting the dataset.

Data Preprocessing

A structured preprocessing pipeline was implemented to prepare the raw Reddit text for clustering. The steps include:

- Removal of HTML tags
- Text normalization and lowercasing
- Removal of URLs and special characters
- Username masking (author_masked) to preserve privacy
- Keyword extraction for each message
- Generation of topic hints

Additional filtering steps were applied to improve data quality:

- Promoted or sponsored posts were removed
- Non-comment permalink entries were filtered
- Low-signal or junk rows were reduced prior to clustering

These steps ensure that the input to the embedding and clustering stages is clean, consistent, and meaningful.

Forum Analysis and Clustering

Embedding (Message Abstraction)

TF-IDF vectorization was used to convert each cleaned message into a fixed-dimensional numerical vector. This transformation represents each post in a high-dimensional feature space based on word importance. The TF-IDF representation satisfies the lab

requirement of embedding or message abstraction prior to clustering.

Clustering

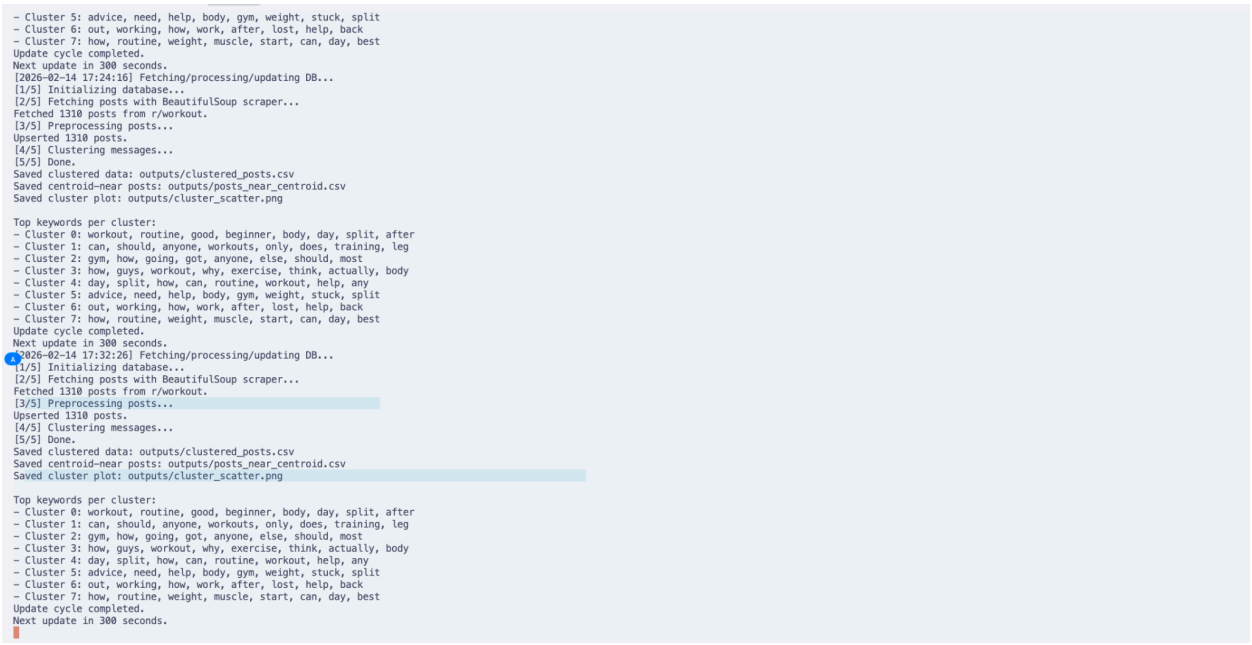
K-Means clustering was applied with K = 8 clusters. The clustering pipeline produces the following outputs:

- Cluster assignment for each message
- Distance from each message to its assigned centroid
- Posts nearest to each centroid
- A two-dimensional cluster visualization generated using PCA projection

Representative Cluster Keywords

From the most recent run, the clusters produced representative keyword sets such as:

Cluster 0: workout, routine, good, beginner, body, day, split, after
Cluster 1: can, should, anyone, workouts, only, does, training, leg
Cluster 2: gym, how, going, got, anyone, else, should, most
Cluster 3: how, guys, workout, why, exercise, think, actually, body
Cluster 4: day, split, how, can, routine, workout, help, any
Cluster 5: advice, need, help, body, gym, weight, stuck, split
Cluster 6: out, working, how, work, after, lost, help, back
Cluster 7: how, routine, weight, muscle, start, can, day, best



These clusters reflect distinct workout-related themes.

Automation

A real-time periodic update system was implemented in automation.py. The script accepts an interval in minutes and executes the following cycle repeatedly:

1. Fetch new posts
2. Preprocess the text
3. Update the database
4. Recompute clustering
5. Wait for the specified interval

During the waiting period, the system accepts free-text queries from the terminal. Each query is embedded and mapped to the nearest cluster. The system then retrieves and displays the most

relevant posts from that cluster. Additionally, a query-specific visualization is generated and saved as a PNG file.

This automation component demonstrates a practical near real-time analysis workflow.

Results and Artifacts

Runtime Summary

In the latest run:

- 500 posts were requested
- 1,277 rows were stored and processed
- 8 clusters were generated

```
Next update in 300 seconds.
back day

Closest cluster: 1
Cluster keywords: can, should, anyone, workouts, only, does, training, leg, split, weight
Top related messages:
- (0) lower back pain is ruining leg day
  https://old.reddit.com/r/workout/comments/1r12dww/lower_back_pain_is_ruining_leg_day/
- (0) What are your back day exercises?
  https://old.reddit.com/r/workout/comments/1qjyg8f/what_are_your_back_day_exercises/
- (0) Slight Back injury, again
  https://old.reddit.com/r/workout/comments/1r40f3d/slight_back_injury_again/
- (0) Gymming @day 1
  https://old.reddit.com/r/workout/comments/1qh2b5c/gymming_day_1/
- (0) Lower back soreness
  https://old.reddit.com/r/workout/comments/1qw8ukg/lower_back_soreness/
/Users/herunkan/GITfolder/DSCI 560/Lab5/automation.py:67: UserWarning: Glyph 128075 (\N{WAVING HAND SIGN}) missing from font(s) DejaVu Sans.
  plt.tight_layout()
/Users/herunkan/GITfolder/DSCI 560/Lab5/automation.py:67: UserWarning: Glyph 128064 (\N{EYES}) missing from font(s) DejaVu Sans.
  plt.tight_layout()
/Users/herunkan/GITfolder/DSCI 560/Lab5/automation.py:68: UserWarning: Glyph 128075 (\N{WAVING HAND SIGN}) missing from font(s) DejaVu Sans.
  plt.savefig(output_path, dpi=160)
/Users/herunkan/GITfolder/DSCI 560/Lab5/automation.py:68: UserWarning: Glyph 128064 (\N{EYES}) missing from font(s) DejaVu Sans.
  plt.savefig(output_path, dpi=160)
Saved cluster view: outputs/query_cluster_1.png

bench press

Closest cluster: 1
Cluster keywords: can, should, anyone, workouts, only, does, training, leg, split, weight
Top related messages:
- (0) Bench press
  https://old.reddit.com/r/workout/comments/1qx7v14/bench_press/
- (0) Dumbbell bench press
  https://old.reddit.com/r/workout/comments/1qxw7wd/dumbbell_bench_press/
- (0) Overcoming fear of barbell bench press?
  https://old.reddit.com/r/workout/comments/1qwyp57/overcoming_fear_of_barbell_bench_press/
- (0) If a workout plan says "bench press into incline bench press" what does that mean exactly?
  https://old.reddit.com/r/workout/comments/1r0mapj/if_a_workout_plan_says_bench_press_into_incline/
- (0) Why use a bench without safeties for bench press if choice available?
  https://old.reddit.com/r/workout/comments/1qu1at3/why_use_a_bench_without_safeties_for_bench_press/
/Users/herunkan/GITfolder/DSCI 560/Lab5/automation.py:67: UserWarning: Glyph 128075 (\N{WAVING HAND SIGN}) missing from font(s) DejaVu Sans.
  plt.tight_layout()
/Users/herunkan/GITfolder/DSCI 560/Lab5/automation.py:67: UserWarning: Glyph 128064 (\N{EYES}) missing from font(s) DejaVu Sans.
  plt.tight_layout()
/Users/herunkan/GITfolder/DSCI 560/Lab5/automation.py:68: UserWarning: Glyph 128075 (\N{WAVING HAND SIGN}) missing from font(s) DejaVu Sans.
  plt.savefig(output_path, dpi=160)
/Users/herunkan/GITfolder/DSCI 560/Lab5/automation.py:68: UserWarning: Glyph 128064 (\N{EYES}) missing from font(s) DejaVu Sans.
  plt.savefig(output_path, dpi=160)
Saved cluster view: outputs/query_cluster_1.png
```

⌘K to generate command

```
mysql> use lab5_workout;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> show tables;
+----------------------+
| Tables_in_lab5_workout |
+----------------------+
| cluster_keywords     |
| reddit_posts         |
+----------------------+
2 rows in set (0.01 sec)

mysql> select * from cluster_keywords limit 10;
+------------+------------------------------------------------------------------+---------------------+
| cluster_id | top_keywords                                                     | updated_at          |
+------------+------------------------------------------------------------------+---------------------+
|          0 | workout,routine,good,beginner,body,day,split,after,help,plan     | 2026-02-14 17:43:37 |
|          1 | can,should,anyone,workouts,only,does,training,leg,split,weight   | 2026-02-14 17:43:37 |
|          2 | gym,how,going,got,anyone,else,should,most,question,people        | 2026-02-14 17:43:37 |
|          3 | how,guys,workout,why,exercise,think,actually,body,out,should     | 2026-02-14 17:43:37 |
|          4 | day,split,how,can,routine,workout,help,any,advice,program        | 2026-02-14 17:43:37 |
|          5 | advice,need,help,body,gym,weight,stuck,split,routine,any         | 2026-02-14 17:43:37 |
|          6 | out,working,how,work,after,lost,help,back,start,advice           | 2026-02-14 17:43:37 |
|          7 | how,routine,weight,muscle,start,can,day,best,body,good           | 2026-02-14 17:43:37 |
+------------+------------------------------------------------------------------+---------------------+
8 rows in set (0.00 sec)

mysql> select * from reddit_posts limit 5;
| post_id | fullname   | created_utc | title                                                                                                    | selftext | clean_text
                                                                  | keywords                                            | topic_hint | author_masked | score | num_comments | url  | permalink
                                                        | is_self | over_18 | stickied | cluster_id | centroid_distance  | fetched_at          |
| 1qe5svj | t3_1qe5svj |  1768536367 | How often do you workout your muscle groups per workout?                                                 |          | How often do you workout your muscle groups per
workout                                                           | workout,how,often,muscle,groups,per                 | workout    | user_185460   |     0 |           15 |      | https://old.reddit.com/r/workout/co
mments/1qe5svj/how_often_do_you_workout_your_muscle_groups_per/  |       1 |       0 |        0 |          3 | 0.9295326859017733 | 2026-02-14 16:47:27 |
| 1qe8boa | t3_1qe8boa |  1768544187 | If you could give one piece of advice to a newcomer in the gym what would it be?                         |          | If you could give one piece of advice to a newco
mer in the gym what would it be                                   | could,give,one,piece,advice,newcomer,gym,would      | could      | user_506388   |     0 |          120 |      | https://old.reddit.com/r/workout/co
mments/1qe8boa/if_you_could_give_one_piece_of_advice_to_a/       |       1 |       0 |        0 |          2 |  0.938189117154896 | 2026-02-14 16:47:27 |
| 1qe9v18 | t3_1qe9v18 |  1768549495 | Fat loss journey                                                                                         |          | Fat loss journey
                                                                  | fat,loss,journey                                    | fat        | user_849930   |     0 |           30 |      | https://old.reddit.com/r/workout/co
mments/1qe9v18/fat_loss_journey/                                  |       1 |       0 |        0 |          1 | 0.9966534823316416 | 2026-02-14 16:47:27 |
| 1qeevda | t3_1qeevda |  1768567138 | The abdomen is the most difficult area for women to develop and tone. I'd also say it's the most difficult to maintain. |          | The abdomen is the most difficult area for women
 to develop and tone I d also say it s the most difficult to maintain | most,difficult,abdomen,area,women,develop,tone,also | most       | user_728583   |     0 |           23 |      | https://old.reddit.com/r/workout/co
mments/1qeevda/the_abdomen_is_the_most_difficult_area_for_women/ |       1 |       0 |        0 |          7 | 0.9699514357067907 | 2026-02-14 16:47:27 |
| 1qef1ju | t3_1qef1ju |  1768567652 | High Reps vs. Heavy Weight: Which actually changed your physique?                                        |          | High Reps vs Heavy Weight Which actually changed
 your physique                                                    | high,reps,heavy,weight,actually,changed,physique    | high       | user_836590   |     0 |          118 |      | https://old.reddit.com/r/workout/co
mments/1qef1ju/high_reps_vs_heavy_weight_which_actually_changed/ |       1 |       0 |        0 |          1 | 0.9923641539551015 | 2026-02-14 16:47:27 |
5 rows in set (0.00 sec)

mysql>
```

⌘K to generate command

Challenges and Decisions

The BeautifulSoup scraping approach was chosen to avoid Reddit API registration requirements.

The subreddit r/Fitness was initially selected but later replaced with r/workout due to excessive repetitive threads naming

Selecting the number of clusters (K = 8) required experimentation to balance interpretability and granularity.

Team Meeting Notes

Day 1 – 2/9/2026

- Reviewed the Lab 5 assignment requirements and clarified deliverables (scraping, preprocessing, clustering, automation, and team notes)
- Discussed scraping options (PRAW vs BeautifulSoup) and agreed to begin with PRAW as the initial approach
- Divided tasks so each team member could focus on scraper setup, database design, and clustering research

Day 2 – 2/10/2026

- Shared progress on Reddit scraping setup and encountered API approval/access limitations with the PRAW route

- Discussed assignment-allowed alternatives and agreed to switch to the BeautifulSoup-based scraping method
- Finalized use of MySQL for storage to align with lab instructions
- Outlined preprocessing requirements including cleaning text, masking usernames, and keyword extraction

Day 3 – 2/11/2026

- Implemented and reviewed the BeautifulSoup scraper with pagination to handle larger post requests
- Built MySQL schema and upsert logic for post storage and cluster metadata
- Integrated preprocessing pipeline and started initial clustering experiments using TF-IDF + KMeans
- Validated that representative posts could be retrieved near each cluster centroid

Day 4 – 2/12/2026

- Developed automation workflow to run scraping, preprocessing, storage, and clustering at fixed intervals
- Added interactive query mode to map user input to the closest cluster and display related messages
- Reviewed output quality and identified irrelevant/promoted/noisy posts affecting cluster quality
- Applied filtering and query-ranking improvements to return more relevant subreddit-specific results

Day 5 – 2/13/2026

- Completed end-to-end testing on r/workout and confirmed successful generation of cluster outputs and query visuals

- Captured terminal evidence/screenshots and verified output artifacts for submission
- Assigned final responsibilities for report formatting, meeting notes export, and repository cleanup