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)

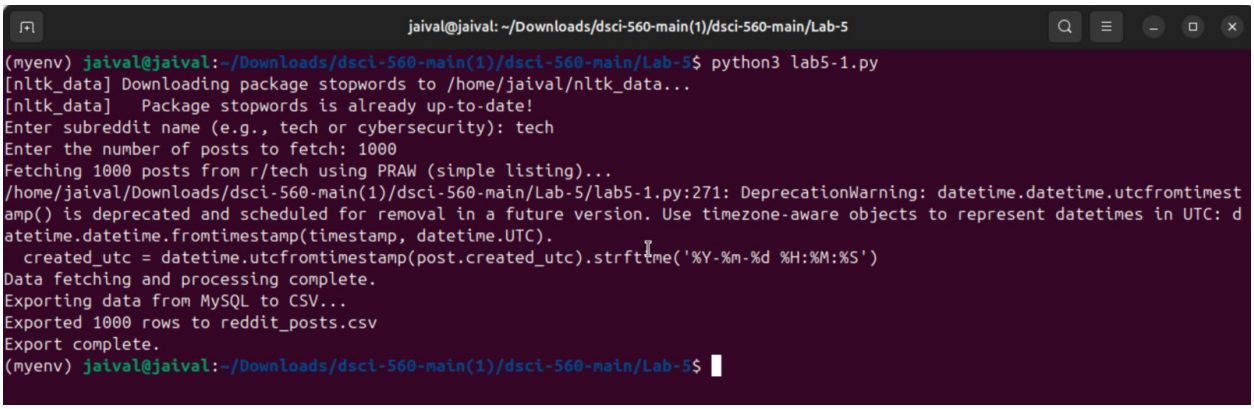
This module is responsible for:

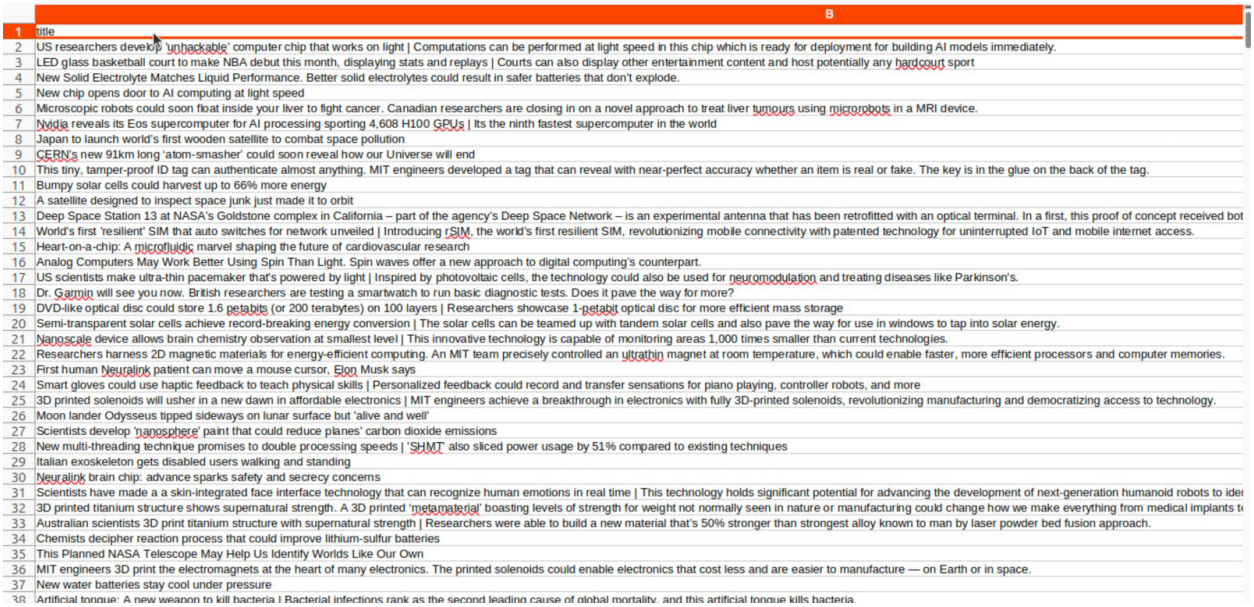
- Scraping Reddit posts using the PRAW API.
- Filtering out irrelevant or promotional posts.
- Cleaning raw text data by removing HTML tags, special characters, and extra spaces.
- Extracting keywords from text using stopword filtering.
- Performing OCR (Optical Character Recognition) on images to extract 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.

Output of Lab5-1.py





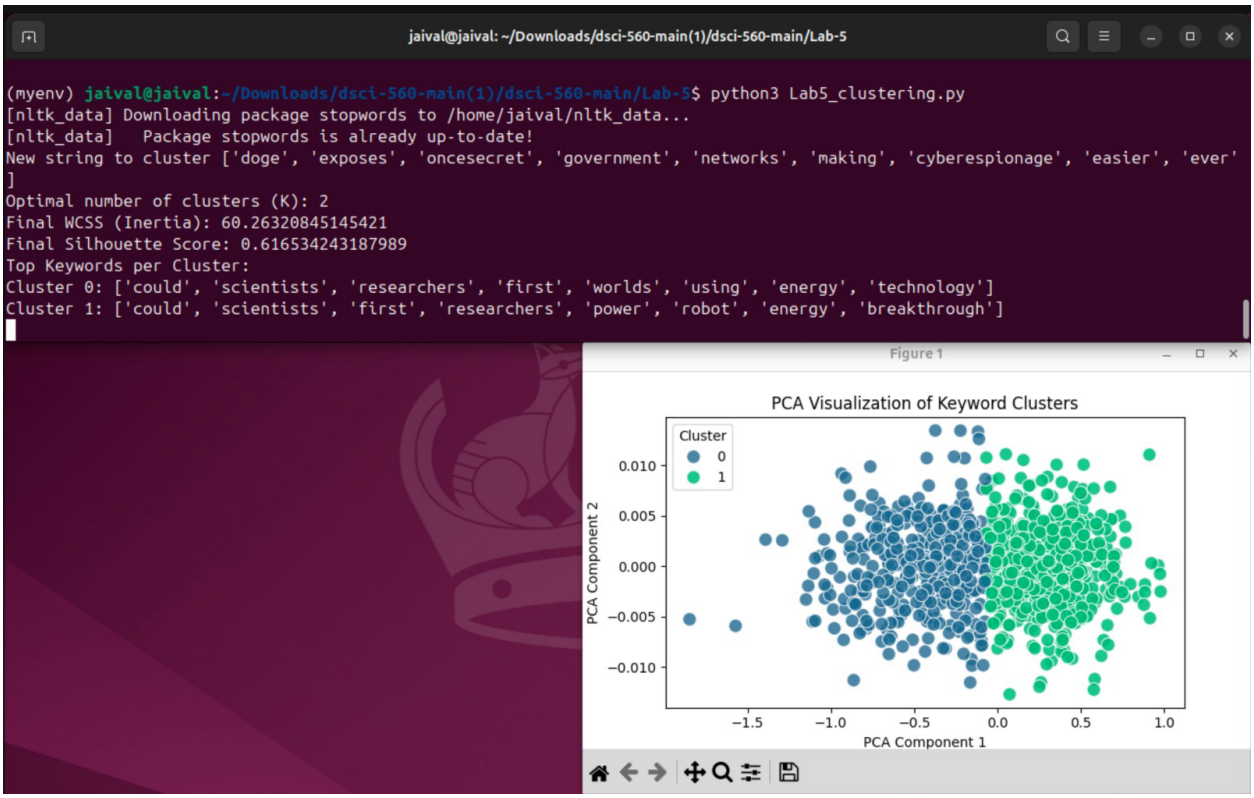
	B
1	title
2	US researchers develop 'unhackable' computer chip that works on light | Computations can be performed at light speed in this chip which is ready for deployment for building AI models immediately.
3	LED glass basketball court to make NBA debut this month, displaying stats and replays | Courts can also display other entertainment content and host potentially any hardcourt sport
4	New Solid Electrolyte Matches Liquid Performance. Better solid electrolytes could result in safer batteries that don't explode.
5	New chip opens door to AI computing at light speed
6	Microscopic robots could soon float inside your liver to fight cancer. Canadian researchers are closing in on a novel approach to treat liver tumours using microrobots in a MRI device.
7	Nvidia reveals its Eos supercomputer for AI processing sporting 4,608 H100 GPUs | Its the ninth fastest supercomputer in the world
8	Japan to launch world's first wooden satellite to combat space pollution
9	CERN's new 91km long 'atom-smasher' could soon reveal how our Universe will end
10	This tiny, tamper-proof ID tag can authenticate almost anything. MIT engineers developed a tag that can reveal with near-perfect accuracy whether an item is real or fake. The key is in the glue on the back of the tag.
11	Bumpy solar cells could harvest up to 66% more energy
12	A satellite designed to inspect space junk just made it to orbit
13	Deep Space Station 13 at NASA's Goldstone complex in California – part of the agency's Deep Space Network – is an experimental antenna that has been retrofitted with an optical terminal. In a first, this proof of concept received bot
14	World's first 'resilient' SIM that auto switches for network unveiled | Introducing rSIM, the world's first resilient SIM, revolutionizing mobile connectivity with patented technology for uninterrupted IoT and mobile internet access.
15	Heart-on-a-chip: A microfluidic marvel shaping the future of cardiovascular research
16	Analog Computers May Work Better Using Spin Than Light. Spin waves offer a new approach to digital computing's counterpart.
17	US scientists make ultra-thin pacemaker that's powered by light | Inspired by photovoltaic cells, the technology could also be used for neuromodulation and treating diseases like Parkinson's.
18	Dr. Garmin will see you now. British researchers are testing a smartwatch to run basic diagnostic tests. Does it pave the way for more?
19	DVD-like optical disc could store 1.6 petabits (or 200 terabytes) on 100 layers | Researchers showcase 1-petabit optical disc for more efficient mass storage
20	Semi-transparent solar cells achieve record-breaking energy conversion | The solar cells can be teamed up with tandem solar cells and also pave the way for use in windows to tap into solar energy.
21	Nanoscale device allows brain chemistry observation at smallest level | This innovative technology is capable of monitoring areas 1,000 times smaller than current technologies.
22	Researchers harness 2D magnetic materials for energy-efficient computing. An MIT team precisely controlled an ultrathin magnet at room temperature, which could enable faster, more efficient processors and computer memories.
23	First human Neuralink patient can move a mouse cursor, Elon Musk says
24	Smart gloves could use haptic feedback to teach physical skills | Personalized feedback could record and transfer sensations for piano playing, controller robots, and more
25	3D printed solenoids will usher in a new dawn in affordable electronics | MIT engineers achieve a breakthrough in electronics with fully 3D-printed solenoids, revolutionizing manufacturing and democratizing access to technology.
26	Moon lander Odysseus tipped sideways on lunar surface but 'alive and well'
27	Scientists develop 'nanosphere' paint that could reduce planes' carbon dioxide emissions
28	New multi-threading technique promises to double processing speeds | 'SHMT' also sliced power usage by 51% compared to existing techniques
29	Italian exoskeleton gets disabled users walking and standing
30	Neuralink brain chip: advance sparks safety and secrecy concerns
31	Scientists have made a a skin-integrated face interface technology that can recognize human emotions in real time | This technology holds significant potential for advancing the development of next-generation humanoid robots to ide
32	3D printed titanium structure shows supernatural strength. A 3D printed 'metamaterial' boasting levels of strength for weight not normally seen in nature or manufacturing could change how we make everything from medical implants t
33	Australian scientists 3D print titanium structure with supernatural strength | Researchers were able to build a new material that's 50% stronger than strongest alloy known to man by laser powder bed fusion approach.
34	Chemists decipher reaction process that could improve lithium-sulfur batteries
35	This Planned NASA Telescope May Help Us Identify Worlds Like Our Own
36	MIT engineers 3D print the electromagnets at the heart of many electronics. The printed solenoids could enable electronics that cost less and are easier to manufacture — on Earth or in space.
37	New water batteries stay cool under pressure
38	Artificial tongue: A new weapon to kill bacteria | Bacterial infections rank as the second leading cause of global mortality, and this artificial tongue kills bacteria.

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.

Output


jaival@jaival: ~/Downloads/dsci-560-main(1)/dsci-560-main/Lab-5
(myenv) jaival@jaival:~/Downloads/dsci-560-main(1)/dsci-560-main/Lab-5$ python3 Lab5_clustering.py
[nltk_data] Downloading package stopwords to /home/jaival/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
New string to cluster ['doge', 'exposes', 'oncesecret', 'government', 'networks', 'making', 'cyberespionage', 'easier', 'ever']
Optimal number of clusters (K): 2
Final WCSS (Inertia): 60.26320845145421
Final Silhouette Score: 0.616534243187989
Top Keywords per Cluster:
Cluster 0: ['could', 'scientists', 'researchers', 'first', 'worlds', 'using', 'energy', 'technology']
Cluster 1: ['could', 'scientists', 'first', 'researchers', 'power', 'robot', 'energy', 'breakthrough']
Figure 1
PCA Visualization of Keyword Clusters
Cluster
0
1
PCA Component 2
PCA Component 1

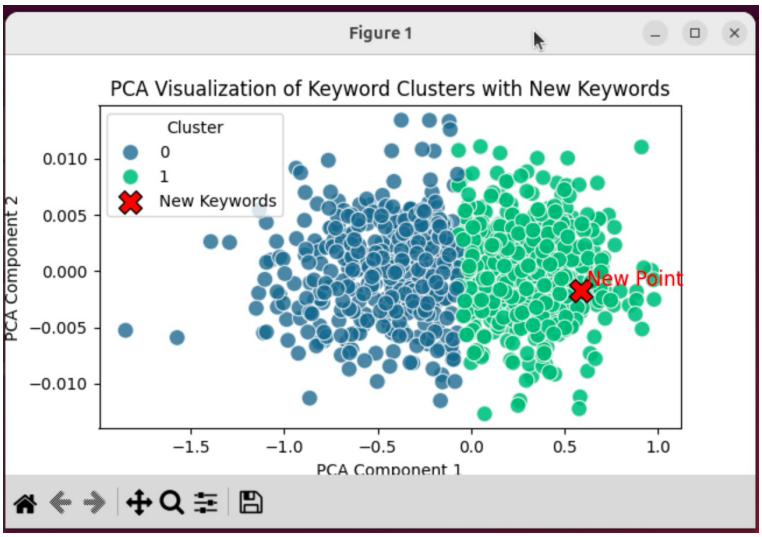
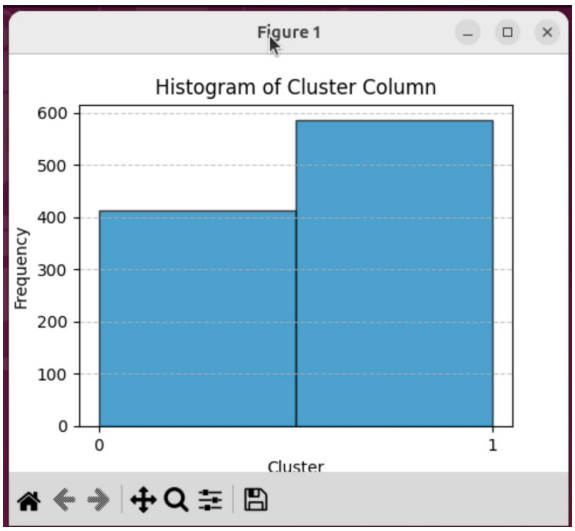
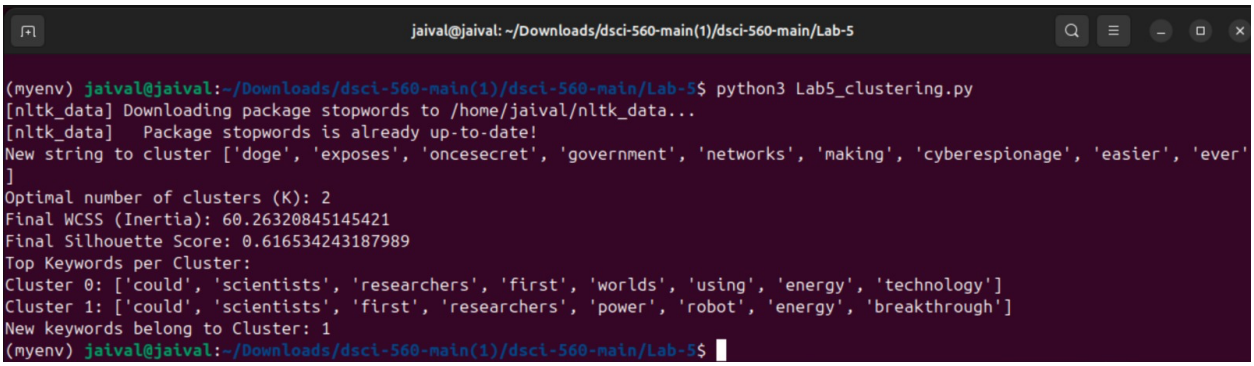


Figure 1
PCA Visualization of Keyword Clusters with New Keywords
Cluster
0
1
New Keywords
New Point
PCA Component 2
PCA Component 1





5) Automation

The automate.py script is designed to automate the execution of two other scripts, scraper.py and clustering.py, at a scheduled interval using the schedule library.

It defines a function run_script(script_name) that:

Executes a given Python script using subprocess.run().

Captures and prints the script's output or errors.

Automates the Workflow

The function automate_process() calls run_script("scraper.py") followed by run_script("clustering.py"), ensuring that data is first scraped and then clustered.

Scheduled Execution

The script takes an interval (in minutes) as a command-line argument.

It schedules automate_process() to run at the specified interval using schedule.every(interval).minutes.do(automate_process).

It enters an infinite loop where it:

Executes any scheduled tasks.

Waits for the next execution, updating the status in real-time.

output

```
(base) bhushanhalasagi@Bhushans-Air Lab-5 % python3 automate.py 1
[INFO] Scheduling automation every 1 minutes.
[INFO] Waiting... Next execution in 1 minutes.
[INFO] Running scraper.py...
[SUCCESS] scraper.py completed successfully.
Fetching 100 posts from r/tech using PRAW (simple listing)...
Data fetching and processing complete.
Exporting data from MySQL to CSV...
Exported 104 rows to reddit_posts.csv
Export complete.


[INFO] Running clustering.py...
[SUCCESS] clustering.py completed successfully.
Optimal number of clusters (K): 2
Final WCSS (Inertia): 0.08724382838874851
Final Silhouette Score: 0.054736801983642044
Top Keywords per Cluster:
Cluster 0: [' researchers', ' scientists', ' cancer', ' cells', ' could', ' patients', ' breakthrough', ' make', ' developed', ' power']
Cluster 1: [' could', ' first', ' treatment', ' robot', ' quantum', ' made', ' milestone', ' world', ' protein', ' soon']
New keywords belong to Cluster: 1

[INFO] Waiting... Next execution in 1 minutes.
[INFO] Running scraper.py...
[SUCCESS] scraper.py completed successfully.
Fetching 100 posts from r/tech using PRAW (simple listing)...
Data fetching and processing complete.
Exporting data from MySQL to CSV...
Exported 104 rows to reddit_posts.csv
Export complete.


[INFO] Running clustering.py...
[SUCCESS] clustering.py completed successfully.
Optimal number of clusters (K): 2
Final WCSS (Inertia): 0.08979684029422448
Final Silhouette Score: 0.06017510039113153
Top Keywords per Cluster:
Cluster 0: [' could', ' first', ' made', ' world', ' protein', ' treatment', ' make', ' robot', ' quantum', ' computer']
Cluster 1: [' researchers', ' cancer', ' scientists', ' could', ' cells', ' breakthrough', ' patients', ' using', ' power', ' help']
New keywords belong to Cluster: 0

[INFO] Waiting... Next execution in 1 minutes.
```

6) Execution Instructions

Prerequisites

Ensure the following dependencies are installed before running the scripts:

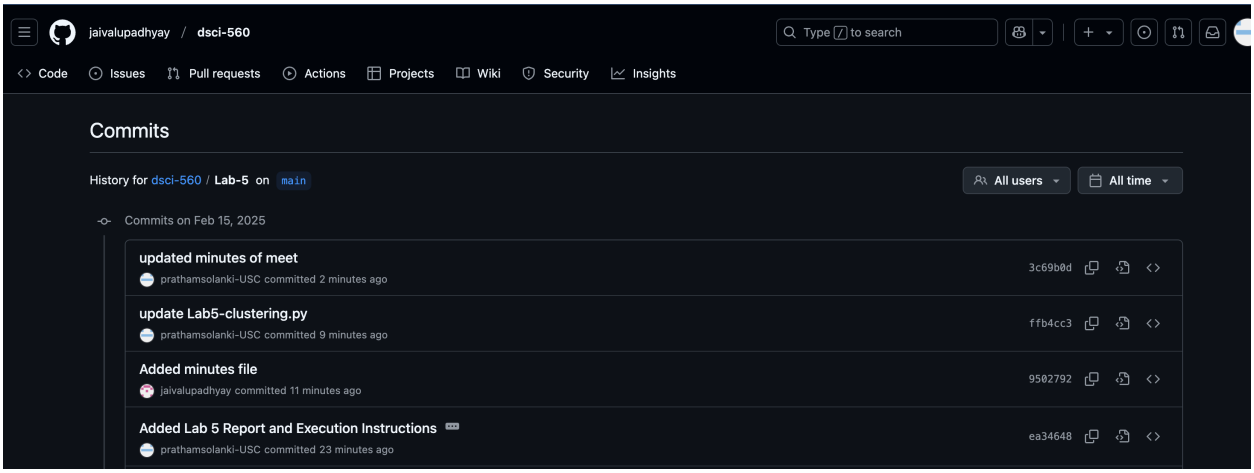
pip install praw mysql-connector-python beautifulsoup4 pytesseract nltk pandas numpy matplotlib seaborn gensim scikit-learn

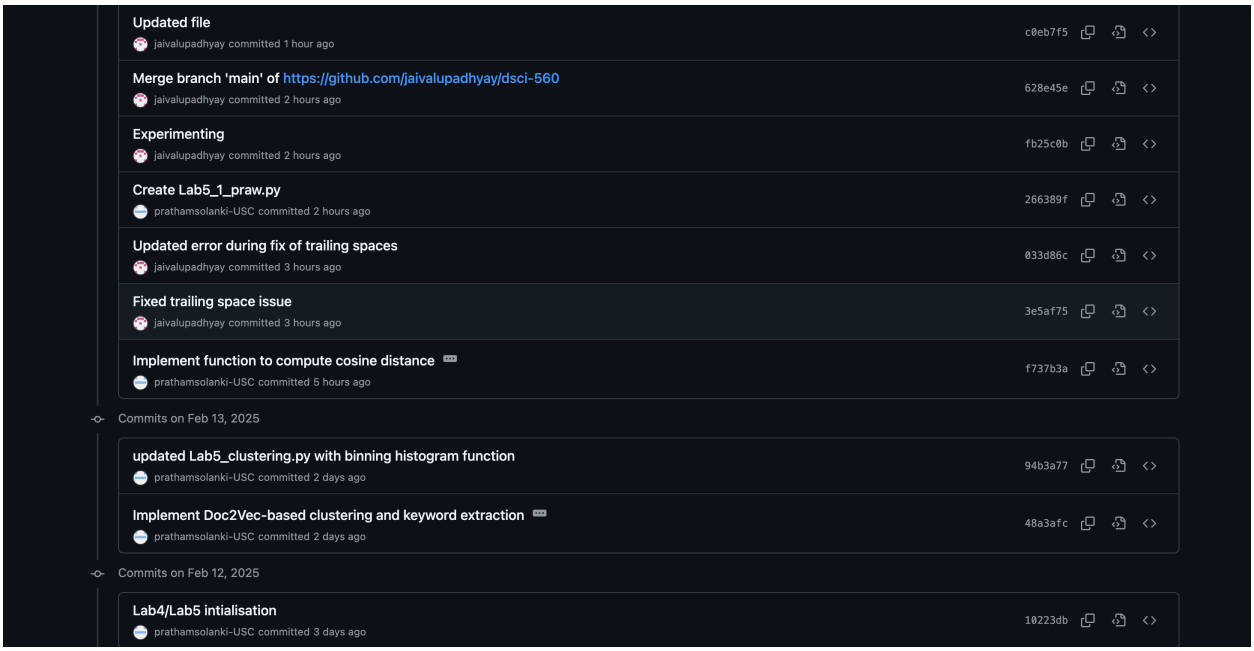
To run
Python3 automate.py

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




Updated file
jaivalupadhyay committed 1 hour ago
c0eb7f5
Merge branch 'main' of https://github.com/jaivalupadhyay/dsci-560
jaivalupadhyay committed 2 hours ago
628e45e
Experimenting
jaivalupadhyay committed 2 hours ago
fb25c0b
Create Lab5_1_praw.py
prathamsolanki-USC committed 2 hours ago
266389f
Updated error during fix of trailing spaces
jaivalupadhyay committed 3 hours ago
033d86c
Fixed trailing space issue
jaivalupadhyay committed 3 hours ago
3e5af75
Implement function to compute cosine distance
prathamsolanki-USC committed 5 hours ago
f737b3a
Commits on Feb 13, 2025
updated Lab5_clustering.py with binning histogram function
prathamsolanki-USC committed 2 days ago
94b3a77
Implement Doc2Vec-based clustering and keyword extraction
prathamsolanki-USC committed 2 days ago
48a3afc
Commits on Feb 12, 2025
Lab4/Lab5 intialisation
prathamsolanki-USC committed 3 days ago
10223db