**Tuesday, February 11, 2025**

**Meeting Agenda:**

* Understanding assignment requirements and objectives.
* Discussing the tools and libraries to be used.
* Assigning initial setup tasks to team members.

**Discussion Points:**

* Decided to use **PRAW** for web scraping as it provides a structured API for Reddit.
* **MySQL** chosen for database storage.
* **BeautifulSoup4** will be used for additional data extraction and preprocessing.
* Discussed automation strategy for periodic data updates.
* Each member will install necessary dependencies and test sample scripts.

**Action Items:**

* [Jaival] Set up PRAW and test Reddit scraping.
* [Jaival] Configure MySQL database and design schema.
* [Pratham] Research OCR tools for extracting text from images.
* [Mayank] Review clustering algorithms for document grouping.

**Wednesday, February 12, 2025**

**Meeting Agenda:**

* Review progress on web scraping and database setup.
* Finalize the Reddit topic selection.
* Plan data preprocessing strategies.

**Discussion Points:**

* Successfully scraped initial set of posts from the selected Reddit topic.
* MySQL schema designed to store messages, timestamps, and keywords.
* Identified necessary preprocessing steps: removing HTML tags, special characters, and filtering irrelevant posts.
* Explored **Pytesseract** for OCR and text extraction from embedded images.
* Decided to use **doc2vec** for text vectorization before clustering.

**Action Items:**

* [Jaival] Implement script for continuous Reddit scraping with API request handling.
* [Jaival] Finalize and test data insertion into MySQL.
* [Pratham] Develop initial preprocessing functions.
* [Pratham] Experiment with doc2vec for text embeddings.

**Thursday, February 13, 2025**

**Meeting Agenda:**

* Integrating preprocessing and database storage.
* Implementing initial clustering algorithm.
* Testing real-time data updates.

**Discussion Points:**

* Web scraping script successfully fetching posts at intervals.
* Preprocessing pipeline now includes text cleaning, timestamp conversion, and username masking.
* **K-means clustering** selected for initial implementation.
* Implemented keyword extraction for cluster labeling.
* Discussed visualization options for displaying clustering results.

**Action Items:**

* [Jaival] Optimize API requests to handle large datasets.
* [Jaival] Test and refine preprocessing script.
* [Pratham] Integrate clustering with database.
* [Mayank] Work on visualization tools for displaying clusters.

**Friday, February 14, 2025**

**Meeting Agenda:**

* Review overall progress.
* Finalize automation script.
* Debugging and testing end-to-end flow.

**Discussion Points:**

* Full pipeline successfully running: **scraping → preprocessing → storage → clustering**.
* Set up an automation script that updates the database every **5 minutes**.
* Implemented a **command-line interface** for querying message clusters.
* Initial visualization showing clusters and keyword distributions.
* Identified areas for improvement: better cluster quality and visualization enhancement.

**Action Items:**

* [Jaival] Improve clustering by experimenting with different embedding techniques.
* [Jaival] Document setup and usage instructions for final submission.
* [Pratham] Refine visualization and user query interface.
* [Mayank] Conduct final testing and debugging