COSC 4P02: Software Engineering II.

Group 9: Progress Report 1

Scrum Master: Jaden Kuhn (jk21pf@brocku.ca - 7249683)

Product Owner: Dea Kukuqani (dk16qs@brocku.ca – 6196018)

Developers: Shijie Tong (st20az@brocku.ca - 7081201)

Nicholas Caruso (yq20oz@brocku.ca - 7189749)

Thomas Semenak (ts19cp@brocku.ca - 6745038)

Dalton Morris (dm20bq@brocku.ca - 7053184)

Chidera Nwana (cn20rq@brocku.ca - 7078124)

Instructor: Naser Ezzati-Jivan

Teaching Assistant: Madeline Janecek

February 23rd, 2025

Table of Contents:

1. Introduction

The AI Powered Newsletter and Social Media Content Generator project aims to make content creation easier and more efficient for newsletters and social media. The goal is to automate tasks like gathering relevant content, summarizing it, and formatting it in a way that is useful for users. By using machine learning, data scraping, and a user-friendly interface, the platform will allow users to generate high-quality content with minimal effort.

This report outlines the progress made in the first three sprints, covering key developments in LLM training, data scraping, and website development. It also details the challenges faced, the decisions made, and how the team is staying on track to meet project deadlines.

2. Sprint Overviews

2.1. Sprint Durations

Sprint 1: January 21, 2025 – January 27, 2025.

Sprint 2: January 28, 2025 – February 17, 2025.

Sprint 3: February 18, 2025 – February 23, 2025.

2.2. Sprint Objectives

The primary objectives for Sprint 1 were:

- Researching possible LLMs and datasets for the project,
- Defining database tables for the project,
- Creating the base login and register pages for later integration with the database.

The primary objectives for Sprint 2 were:

- Finalizing and training the Large Language Model (LLM) for content generation,
- Developing a robust data scraping mechanism for content aggregation.

The primary objectives for Sprint 3 were:

- Improving the SERP API data scraping tool and developing a HTML scraping tool to obtain the contents from each URL,
- Hosting the application on Azure,
- Enhancing the user interface and integrating backend login functionalities.

3. Contributions and Key Decisions

3.1. Large Language Model (LLM) Training

Team Members Involved: Jaden, Nicholas, Dea.

Objectives:

- Identify and finalize a Large Language Model (LLM) suitable for content generation,
- Train and optimize the model for efficient and high-quality text generation.

Decisions and Progress:

- **Model Selection:** The team evaluated multiple LLMs, including OpenAI's GPT models and open-source alternatives. Deepseek was selected due to its balance of cost-efficiency, reliability, and performance.
- **Fine-Tuning:** Initial testing demonstrated that Deepseek produced high-quality outputs without requiring extensive fine-tuning. This decision helps reduce computational costs and facilitates faster model deployment.
- **Training Scripts:** Training scripts were developed and uploaded to GitHub to ensure consistent content generation tailored to user preferences.

Impact on Project Timeline:

- The finalized LLM allows the team to proceed with integration rather than continued research.
- Training optimizations reduce system overhead, making the model cost-effective for deployment.

3.2. Data Scraping Development

Team Members Involved: Thomas, Dalton, Shijie, Jaden.

Objectives:

- Implement a data scraping module to retrieve and aggregate relevant content (URLs),
- Ensure flexibility and accuracy in content filtering,
- Implement a web scraping algorithm to retrieve the HTML content from the aggregated URLs.

Decisions and Progress:

- **Initial API Testing:** The team initially integrated News API for content retrieval. However, testing revealed limitations in search functionality and unreliable content filtering.
- **Alternative API Integration:** Due to the limitations of News API, the team decided to switch to SERP API instead, a far superior API. A functional integration prototype was uploaded to GitHub that:
  - Uses topic-based searching to retrieve articles.
  - Provides enhanced filtering for improved accuracy and relevance.
- **Current Status:** The SERP API integration is partially complete, with additional search parameters being developed. Manual testing is underway to refine search accuracy.
- **BeautifulSoup API:** For the web scraping algorithm, we decided to use the BeautifulSoup API to gather the HTML content. Currently, we have the algorithm mostly functional as it scrapes the HTML content of specific tags and filters it further to reduce noise. It is complete as of now, but extra work could be done to process the returned content more.

Impact on Project Timeline:

- Moving away from less capable APIs reduces reliance on external services, ensuring project stability.
- The new API integration allows parallel development with other project components.
- The completion of the web scraping (HTML) content allows us to now integrate that algorithm with the LLM.

3.3. Website Development and Integration

Team Members Involved: Chidera, Dea.

Objectives:

- Improve frontend user interface and integrate backend functionalities,
- Implement a secure authentication system.

Decisions and Progress:

- **User Interface Development:**
  - The landing page was redesigned for improved navigation.
  - The dashboard was updated based on user feedback to enhance user experience.
- **Authentication System:**
  - The login and registration system were finalized and connected to the database.
  - Security enhancements were implemented to prevent unauthorized access.
- **Backend Integration:**
  - Work is ongoing to ensure seamless connectivity between the frontend user interface and backend services.

Impact on Project Timeline:

- A fully functional authentication system enables real user testing.
- Finalized user interface components prevent major redesigns in future sprints.

3.4. Website Hosting

Team Member Involved: Jaden.

Objectives:

- Find a suitable service to host our website and backend,
- Get some domain and host our website so anyone can access it.

Decisions and Progress:

- **Azure:** We decided to go with Azure as some of us were familiar with it and Azure provides free credit to use within the month. Azure allows us to host not only our website, but also our DB and backend code.
- **Current Status:** Our website is hosted and can be accessed by anyone, anywhere. Our next steps are to set up the database environment and integrate that with the website.

Impact on Project Timeline:

- A hosted application enables a real-time view of the application after every change.

- Significant time will be saved in future sprints with the application already hosted.

4. Challenges

4.1. Large Language Model (LLM)

During the first sprint, the team researched different LLMs and datasets to see what would work best for our project. By the end of the sprint, we decided to use T5, and began the implementation and training in second sprint. We then discovered that this base model was not very good, and the datasets we chose would not work as it was mainly for unsupervised learning. As a result, we needed to change our LLM and datasets in the middle of the second sprint, which put us behind temporarily. We were able to successfully find a superior model (Deepseek) and were able to get back on track for the remainder of the project.

4.2. Content (URL) Aggregation

The primary challenge we encountered was the limited scope of the News API, as its source list was not expansive enough and it only allowed scraping a limited amount of historical data. In response, we explored custom web scraping in Sprint 2 but encountered technical difficulties that made it a challenging and time-consuming solution. This was a significant hurdle for the team, particularly during a busy school week with other coursework adding to the pressure. This obstacle emerged as our group’s main challenge, as significant effort was invested with minimal progress. However, the team responded in the following week committing more time and effort to overcome this challenge.

4.3. Website Development

One of the challenges faced with development of this website was the struggle with structuring the layout, particularly with button placement and implementation of the tag functionality. These issues caused problems the overall usability and user experience of the dashboard. There was also a challenge on getting the sign up and login button to lead to the correct pages while maintaining the functionalities. However, through collaboration with Dea, we were able to address these problems. By rearranging major aspects of the dashboard and fixing major issues with the button placements and functionality, a more user-friendly and intuitive interface was created, and most buttons will be active and lead to the correct pages before next stage.

5. Sprint 1+2+3 Task Breakdown (Jira Board Updates)

Task	Status	Team Members
LLM Research and Selection	Completed	Jaden, Nicholas, Dea
LLM Training Scripts Development	Completed	Jaden, Nicholas, Dea
Initial Data Scraping Module (News API)	Completed	Thomas, Dalton, Shijie
Enhanced API integration	In Progress	Thomas, Dalton, Shijie
Web Scraping (HTML content) algorithm development	Completed	Jaden
Frontend UI Updates (Landing Page, Dashboard)	Completed	Chidera, Dea
Authentication System Implementation	Completed	Chidera, Dea
Backend Integration with UI	In Progress	Chidera, Dea
Azure website integration	Completed	Jaden
Testing and Debugging	In Progress	All Members

6. Overall Sprint Assessment and Next Steps

6.1 Summary of Sprint 1 Achievements

- Researched different LLMs and datasets to make an informed decision on a model,
- Defined database tables for the login/registration system, as well as for users,
- Created the login and registration pages once the database tables were created. Integration will be done in a future sprint.

6.2 Summary of Sprint 2 Achievements

- Finalized Deepseek as the LLM model and optimized it for content generation,
- Developed a functional data scraper with improved content filtering,
- Implemented a secure authentication system for users.

6.3 Summary of Sprint 3 Achievements

- Hosted our website on Azure,
- Redesigned landing page and user dashboard, enhancing the user interface and user experience,
- Developed a web scraping algorithm to gather the HTML content of a given URL.

6.4 Remaining Priorities for Next Sprints

- Complete integration of SERPAPI to replace News API.
    - Additional Features:
        - Analytics: Implement real-time analytics with dynamic graphs that track changes in keyword interest over time, allowing users to monitor trending topics.
        - User history: Introduced a feature that records user history, enabling personalized article suggestions and the ability to save favorites.
        - Expand the platform by incorporating additional APIs, including those from social media platforms (X, Reddit, Facebook) and blog sources, to diversify data inputs.
- Finalize backend integration to connect all platform components.
- Conduct comprehensive testing to identify and resolve potential issues.
- Begin trial deployment and user testing for feedback.

7. Conclusion

Sprint 1 was a short sprint heavily focused on research and the initial setup of the project. The LLM team focused on finding a model that would work well with our goals for the project while considering the time and budget available. The database team setup the required tables for the project for future integration, and the website team created the login and registration pages of the application. The team easily maintained the project timeline through this sprint and focused on the next sprint where the LLM was finalized and trained, data scraping APIs were used and tested, and the main website underwent further updates.

Sprint 2 was completed successfully, with significant progress in model training, data scraping, and user interface development. The LLM model has been finalized and trained, the website has been transformed, and a functional API has been located and implemented for data scraping. The team has maintained the project timeline by making informed decisions and proactively addressing challenges. The next sprint focused on finalizing backend integration, completing the SERP API integration, and preparing the system for testing and deployment.

Sprint 3 was also completed successfully, with our website being hosted on Azure, the landing page and dashboard being redesigned, SERP API being improved, and a web scraping algorithm being implemented and tested on various URLs. The next sprint will focus on integrating the various aspects of the previous sprints, including the web and data scrapers with the LLM, and the login/authentication system with the database. Regular updates will continue to ensure the project remains on schedule through the final sprints.

Appendix A. Website and GitHub Repository Links

Website Link: https://group9test-ese2cvbrhed8dsfp.canadaeast-01.azurewebsites.net/

GitHub Repository Link: https://github.com/JadenKBrock/4P02GroupProject

Appendix B. Meeting Minutes

Meeting 4 – January 21, 2025

Attendees: Dea, Dalton, Jaden, Nicholas, Thomas, Chidera, Shijie

- Discussed Sprint 1 task distribution.
- Teams assigned for research, UI development, and database setup.
- Communication was set to take place via Microsoft Teams.

Meeting 5 – January 28, 2025

Attendees: Dea, Dalton, Jaden, Nicholas, Thomas, Chidera
Absent: Shijie

- Sprint 1 review confirmed LLM research completion, UI implementation, and database setup.
- Sprint 2 tasks were allocated to teams.

Meeting 6 – February 4, 2025

Attendees: Dea, Dalton, Jaden, Nicholas, Thomas, Chidera
Absent: Shijie

- Mid-Sprint 2 progress review.
- LLM narrowed to GPT and Deepseek, with training initiated.
- Scraper prototype uploaded to GitHub.

Meeting 7 – February 11, 2025

Attendees: Dea, Jaden, Nicholas, Thomas, Shijie
Absent: Dalton, Chidera

- Deepseek selected as the LLM model.
- Landing page uploaded to GitHub.
- Identified issues with NewsApi and initiated the transition to SERPAPI integration

Meeting 8 – February 17, 2025

Attendees: Dea, Dalton, Jaden, Nicholas, Thomas, Chidera, Shijie

- Confirmed all teams were on track for Sprint 2 completion.
- Data scraper demonstrated improved results.
- Dashboard redesign finalized and uploaded.

Meeting 9 – February 18, 2025

Attendees: Dea, Dalton, Jaden, Nicholas, Thomas, Chidera, Shijie, Madeline, Naser.

- Presented the current state of our project to Madeline for feedback and verification.

Appendix C. Complete Contributions Table:

Team Member	**Contribution**
Jaden Kuhn	Sprint 1: - Researched multiple different possible LLMs and datasets for the next sprint Sprint 2/3: - Ran tests with the LLM we picked from Sprint 1 and discovered that

	both the LLM and datasets were probably not going to work. - Ran some more tests with the LLM Deepseek (which we ended up choosing for our LLM) - After we chose to stick with Deepseek, I experimented with changing the model parameters and prompt, and eventually got good results from the model - Created an Azure web app along with a resource group to host everything in our project - Connected the GitHub repository with Azure and got the site working (hosted and running) - Created a web scraping algorithm which takes a URL and scrapes the content of the website
Chidera Nwana	Sprint 1: - Developed tables and queries to be used for database management for the website using MySQL Sprint 2/3: - Created the initial draft of the dashboard for the website - Implemented initial sorting functionality to allow content to be sorted by popularity, date added and tags - Collaborated with Dea to cleanup initial issues with the dashboard, implement better button arrangement and reorganize the UI for enhance use and easier navigation.
Dalton Morris	Sprint 1:

	- Helped with implementing login and registration page and tidying it up in terms of the backend code and UI Sprint 2/3: - Added parsing error checking for articles for the original web scraping API, so when an article object was created it would check for bad/incorrect URLs and other errors. - Added more exception handling for the newest webscraping API. - Touched up the history storing and display for the newest webscraping API. - Touched up code (i.e. mitigated the need for a hardcoded API key).
Dea Kukuqani	Sprint 1: - Focused on researching possible LLMs and Datasets. Sprint 2/3: - Continued on testing different mixes of the LLMs and Data sets before making a final decision with the rest of the group. - Tested out different prompts to be used as well in order to achieve the goal of summarizing content while maintaining the integrity of the article as well as making sure it meets the criteria of different social media platforms - Collaborated with Chidera to get the main landing page working and focused on UX/UI research to make sure that users can achieve their goals on our site and get access to different features through the use of

	different active buttons and simultaneously making sure the design and aesthetic of our site creates a good experience for the user.
Nicholas Caruso	Sprint 1: - Researched various LLMs and datasets for the project, including models like Pegasus, GPT, BLOOM, and datasets like Newsroom, Reddit, XSum. Sprint 2/3: - Ran initial tests with the T5 model locally. - Ran tests with the Deepseek model locally, which became our model. - Experimented with Deepseek locally, and using different platforms such as Colab and Hugging Face Spaces.
Thomas Semenak	Sprint 1: - Developed a local database to support user registration and login Sprint 2/3: - Experimented with a custom web scraping solution using the Beautiful soup library as an alternative to the News API - Collaborated with Shijie to evaluate alternative APIs, ultimately identifying Serpapi as the best option. - Combined keywords from both google_search and google_news into a single result set, offering users enhanced flexibility with what

	appeared to be the optimal parameters - Built a template HTML file to prototype how search results would be displayed on a webpage.
Shijie Tong	Sprint 1: – Finished the login and registration web page (connect with SQL, most of code, logic part) – Improved webpage style Sprint 2/3: – Helped work with News API – Wrote a web scraper function for this first API – Found a new API and worked with it – Wrote a history function for new API – Fix SQL code, link new api apps and login and registration to each other. – All the above in my part already tested in my side. – It starts working: Users can create their own SQL tables to store their search history when they sign up. Users then log in to link their own form to facilitate later project structure. But there are still some problems with logging in and going to the app.

Appendix D. Jira Sprint Burndown Charts

Sprint 1:

After consulting with Madeline towards the end of sprint 1, we learned that story point estimates were needed for each user story and subtask in Jira. Despite them being added, Jira does not update this on an open sprint (the story points remained 0 even though they were added before the end of the sprint). Therefore, we do not have a complete sprint burndown chart for sprint 1. The “Remaining Work” and “Guideline” lines were flat for this sprint. The intended work for the sprint was still completed on time, and this was corrected for future sprints.

Sprint 2:

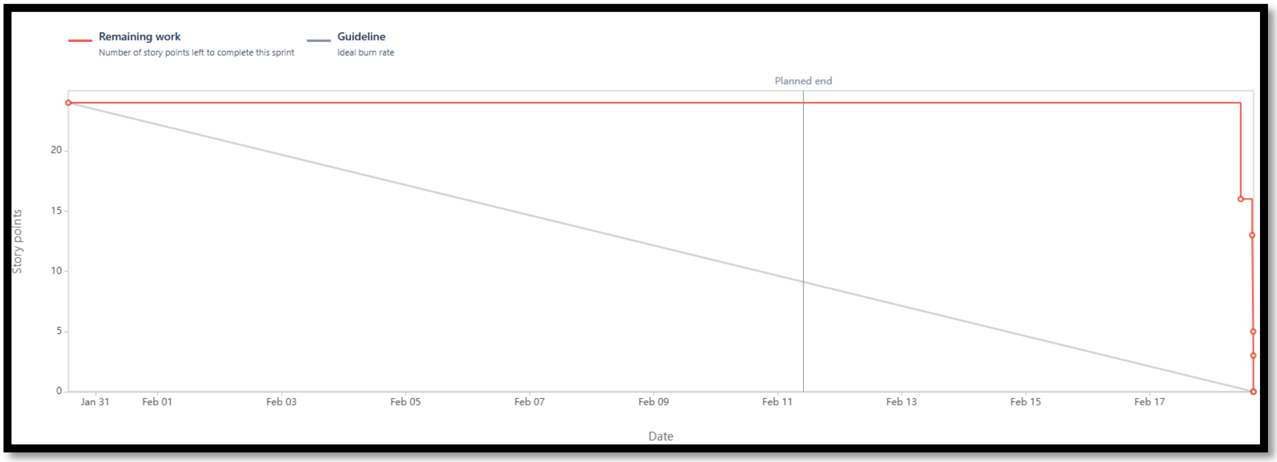


*Figure 1: Burndown Chart for Sprint 2. * See Note Below **

Note: For this sprint (2), we believed that marking subtasks as complete would show the completed story points in the burndown chart. We discovered late in the sprint after only marking subtasks as complete that only points associated with user stories are considered for the burndown chart. This is the reason that the "Remaining Work" line is mostly horizontal. The required work for sprint two was completed throughout the entire sprint, not just at the end as the graph indicates.

Sprint 3:

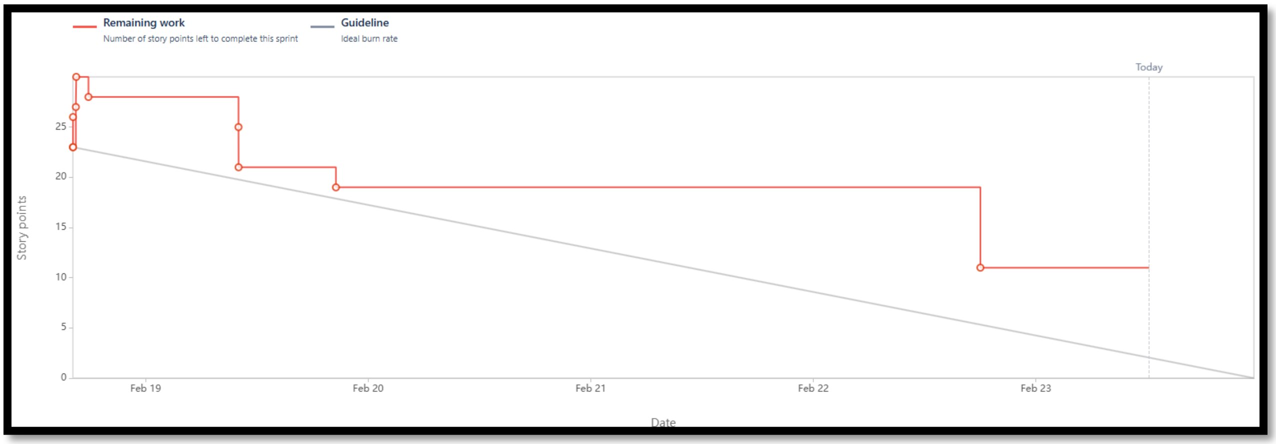
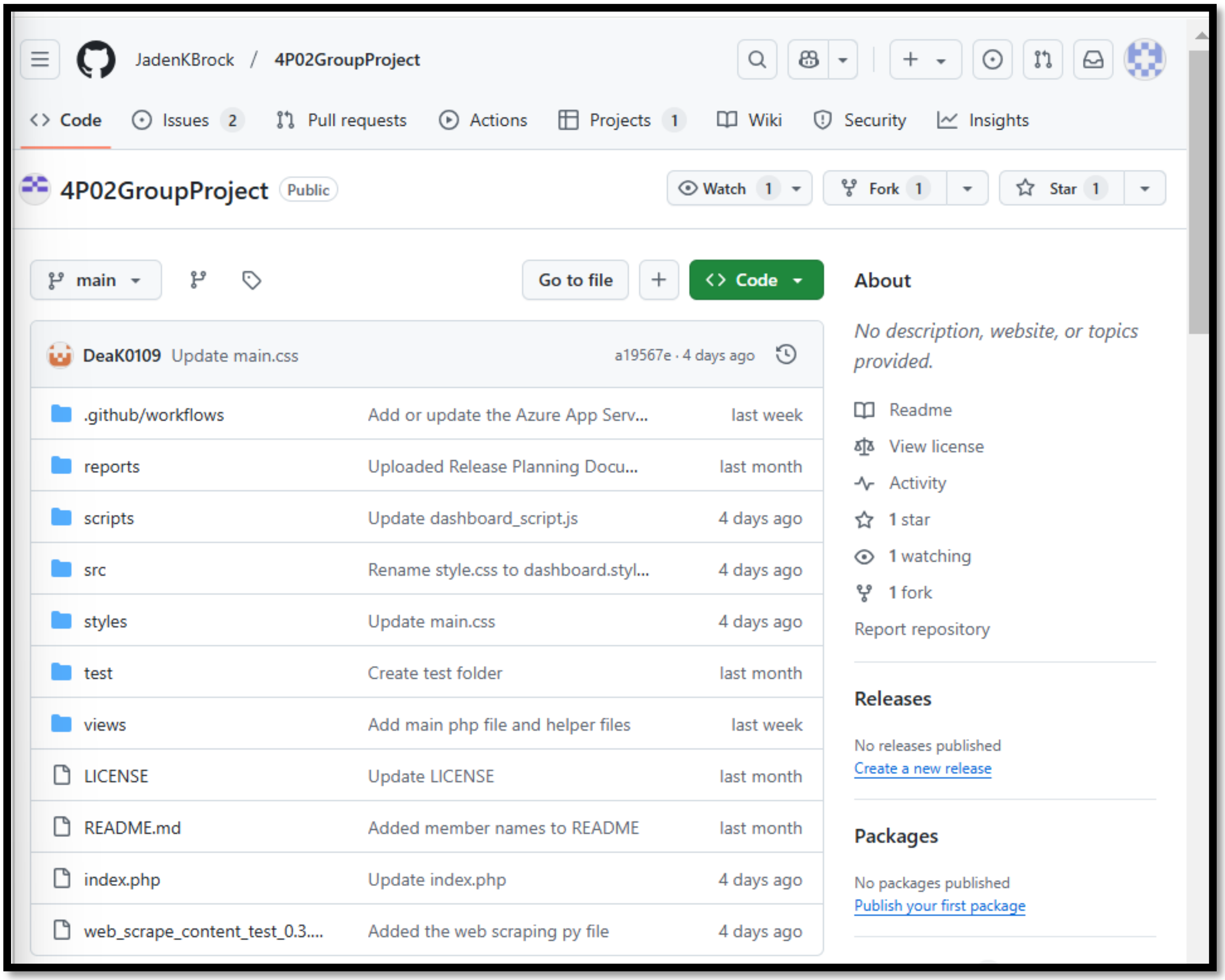


Figure 2: Burndown Chart for Sprint 3.

Note: This sprint officially ends tonight (screenshot taken before the end of the sprint), so the remaining tasks will be marked done at the official end of the sprint for discussion in the retrospective meeting on Tuesday.

Appendix E. GitHub Screenshots



DeaK0109 Update main.css a19567e · 4 days ago

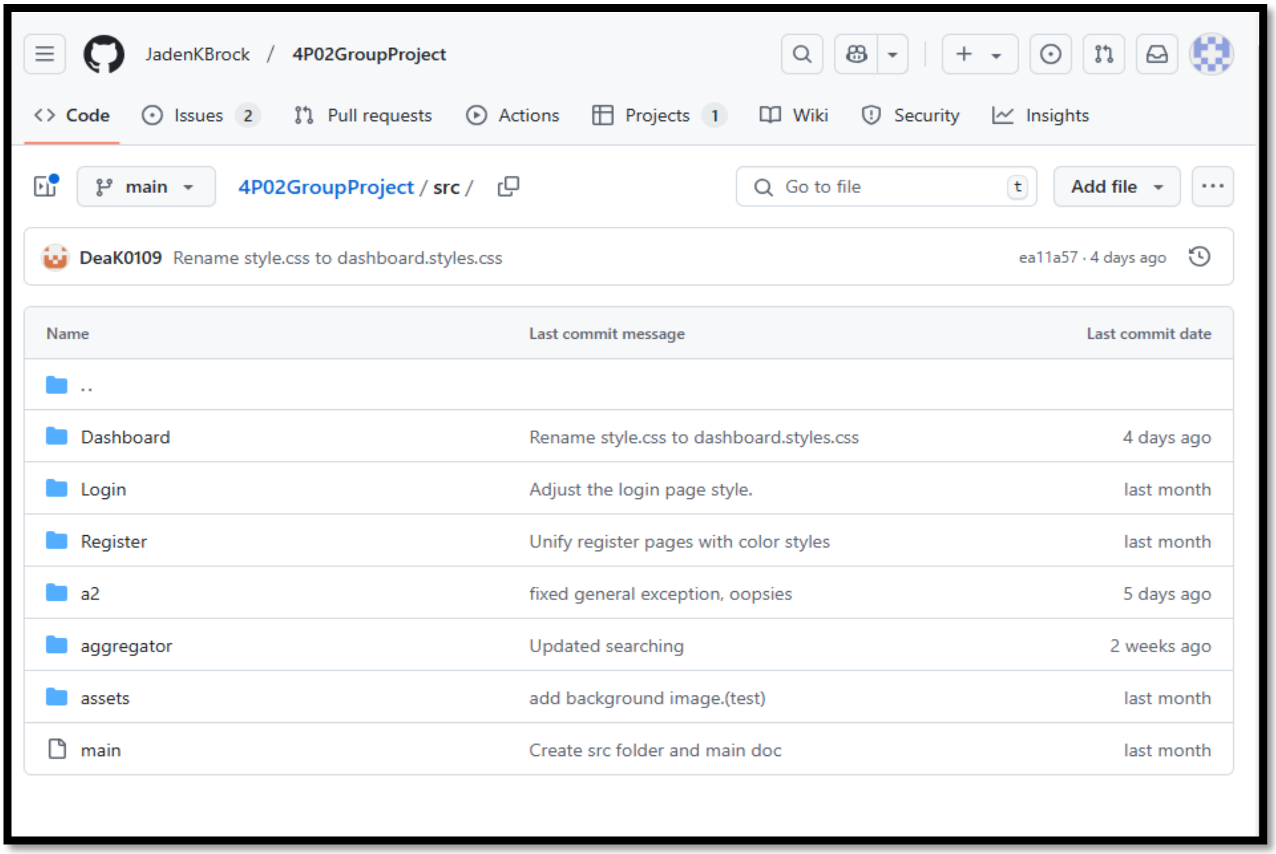
.github/workflows	Add or update the Azure App Serv...	last week
reports	Uploaded Release Planning Docu...	last month
scripts	Update dashboard_script.js	4 days ago
src	Rename style.css to dashboard.styl...	4 days ago
styles	Update main.css	4 days ago
test	Create test folder	last month
views	Add main php file and helper files	last week
LICENSE	Update LICENSE	last month
README.md	Added member names to README	last month
index.php	Update index.php	4 days ago
web_scrape_content_test_0.3....	Added the web scraping py file	4 days ago

No description, website, or topics provided.

No releases published

No packages published

Figure 3: The Main Branch of our GitHub Repository.

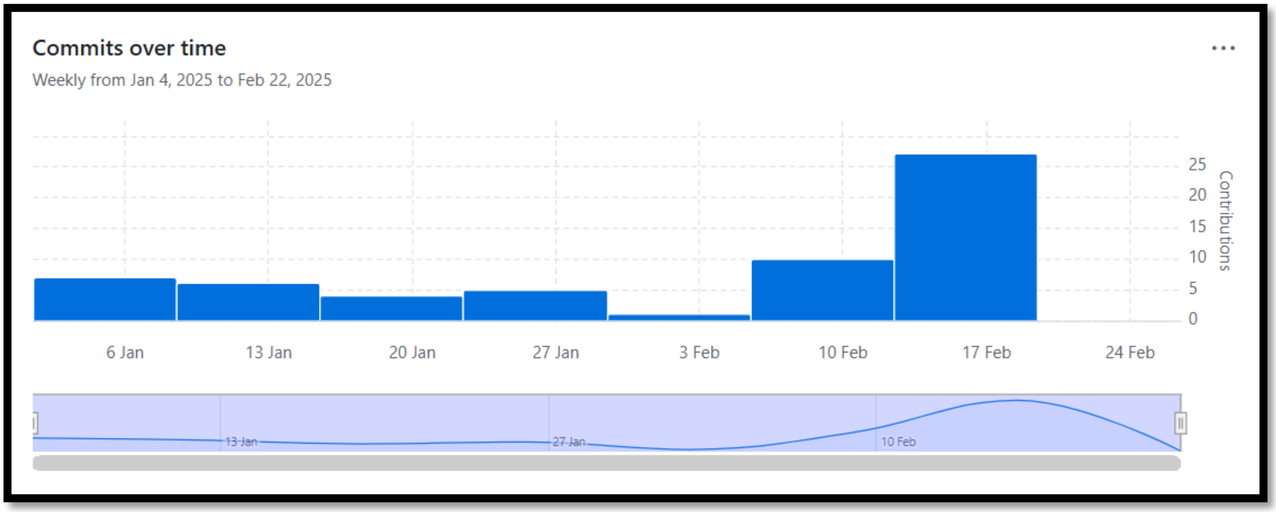


4P02GroupProject / src /

DeaK0109 Rename style.css to dashboard.styles.css ea11a57 · 4 days ago

Name	Last commit message	Last commit date
..		
Dashboard	Rename style.css to dashboard.styles.css	4 days ago
Login	Adjust the login page style.	last month
Register	Unify register pages with color styles	last month
a2	fixed general exception, oopsies	5 days ago
aggregator	Updated searching	2 weeks ago
assets	add background image.(test)	last month
main	Create src folder and main doc	last month

Figure 4: The Source Folder of our GitHub Repository.


Commits over time

Weekly from Jan 4, 2025 to Feb 22, 2025

Figure 5: The "Commits Over Time" Graph.