**Reddit Insights Chatbot with RAG**

**MSDS692 – Data Science Practicum I**  
Weekly Progress Reports  
**Jonish Bishwakarma**

**Week 3 – Data Cleaning and Preprocessing**

This week my focus was on cleaning and filtering the dataset we collected from Reddit. After running the updated scraping code, the dataset grew to around 29,000 rows. However, not all this information was relevant to our project goals, since many posts and comments did not mention any software or tools. To make the dataset more useful, I worked on preprocessing it and filtering down to only the rows that contained mentions of industry-specific software.

To do this, I merged post titles and content into a single column called full\_text and created dictionaries of keywords for each category (Law, Construction, and Tech). These dictionaries included common tools and software used in each industry, like **Clio, Westlaw, Filevine, Bluebeam, Procore, AutoCAD, Jira, Linux, AWS, Firewalls, etc**. Using these keyword lists, I flagged posts and comments that mentioned at least one tool. This filtering process reduced the dataset significantly, from 29,077 total rows to 696 relevant rows, which is much more manageable for analysis.

I also cleaned the text by removing links, special characters, and extra spaces. This will make it easier to prepare for embeddings later when we work on the RAG chatbot. With the filtered dataset, I started doing exploratory analysis to get an initial sense of the conversations. Most of the mentions came from the Tech category (577 mentions), followed by Law (113 mentions), and only a few in Construction (6 mentions). This suggests that tech-related subreddits are currently much more active in discussing tools and software.

Looking at specific tools, Linux was the most mentioned overall (349 mentions), followed by Firewalls (75 mentions), Azure (53 mentions), and Clio (42 mentions). This aligns well with our project since the local business partner is interested in tracking discussions about different kinds of industry software. The subreddit breakdown showed that most tool mentions came from **r/linux, r/sysadmin, r/LawFirm, r/legaltech, and r/msp**. These subreddits might be particularly valuable to monitor moving forward.

Overall, this week I was able to transform the large and noisy dataset into a cleaner, focused dataset that highlights the conversations relevant to our end goal. This feels like an important step because now we can start drawing meaningful insights from the data rather than being overwhelmed by random Reddit posts.

**Project Timeline**

✅Week 1 – Project proposal submission

✅ Week 2 – Data collection from Reddit

✅ Week 3 – Data cleaning, preprocessing, and keyword filtering

⬜ Week 4 – Build software dictionary and start extracting mentions

⬜ Week 5 – Sentiment analysis and identification of pain points

⬜ Week 6 – Develop and test RAG + LLM pipelines

⬜ Week 7 – Evaluation of RAG vs LLM and refinement

⬜ Week 8 – Final analysis, presentation preparation, and submission

**Exploratory analysis**

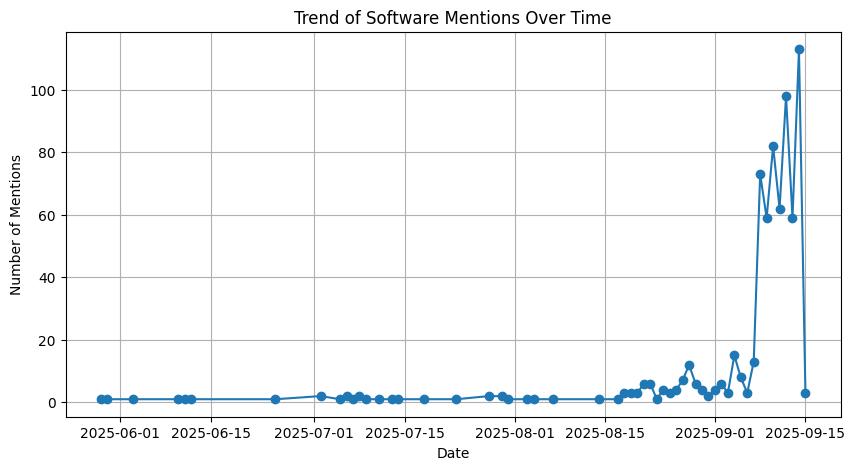
A graph of a bar chart

AI-generated content may be incorrect. A bar graph with different colored squares

AI-generated content may be incorrect.

A graph with blue and green bars

AI-generated content may be incorrect.



**Plan for Next Week**

For Week 4, my plan is to expand our dataset by adding new posts and comments and the software dictionaries further by adding more relevant tools that may not yet be included. I will also refine the filtering process to capture broader mentions while avoiding noise. At the same time, I will start summarizing the context in which these tools are being discussed (for example, positive vs. negative comments, feature requests, or complaints). This will move us closer to identifying pain points in each industry, which is an important step before building the RAG chatbot.