**Reddit Insights Chatbot with RAG**

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

**Week 2 – Data Collection**

This week my focus was on data collection from Reddit. Since the project involves analyzing industry-specific conversations, I chose two subreddits to start with: /*Construction* and /*legaladvice*. These match the domains we talked about with the business partner (construction and law firms).

I wrote Python scripts using the Reddit API (via PRAW) to pull both posts and comments into a CSV format. At first, it was a little confusing how to structure the data because each post can have multiple comments and replies. Eventually, I decided to store every post, comment, and reply as its own row, with a column that links comments back to their parent post. This way we won’t lose the discussion context, and it should be manageable later when we do analysis.

The dataset now includes fields like post ID, subreddit, title, content, author, score, number of comments, created time, and whether it was edited. At first, I didn’t understand why the “title” column was blank in most rows, but then I realized it’s because comments don’t have titles, only posts do. So, the structure makes sense now. It looks a bit complicated in Excel since there are many rows for each discussion thread, but I think this format will work for extracting mentions of software and pain points later.

I haven’t yet added any keyword filtering for specific tools or software, but that will be the next step. The goal is to identify tools like AutoCAD, Procore, Clio, or Westlaw depending on the industry, and then track what people are saying about them. For now, I just wanted to make sure we could reliably pull the raw Reddit data and store it properly.

So far, I don’t see any major roadblocks, but one question I had was about pulling edits and thread history. Reddit marks if something has been edited, which I am capturing, but I don’t think I’m saving the full edit history. That might not be possible with the API, but I’ll look into it more.

Overall, this week I set up the data collection pipeline and have working CSV files for the Construction subreddit. I’ll expand this to /legaladvice next and then start building the dictionaries of industry-specific software so we can begin extracting insights in the coming weeks.

**Project Timeline**

✅Week 1 – Project proposal submission

✅ Week 2 – Data collection from Reddit (Construction + legaladvice)

⬜ 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

**Plan for Next Week**

Next week, my plan is to start cleaning and preprocessing the collected data. This will include removing duplicates, stripping out deleted or removed content, and preparing text for analysis. I will also begin experimenting with keyword searches to detect software names relevant to each industry, such as project management tools in construction or case management software in law. The goal for Week 3 is to move beyond raw collection and start shaping the dataset into something usable for insights.