**Reddit Insights Chatbot with RAG  
MSDS692 – Data Science Practicum 1**

**Jonish Bishwakarma**

**Progress Report for Week4**

**Project Details**This project aims to build a research tool that uses Reddit discussions to surface industry-specific insights, particularly about commonly used software and related pain points in law firms, construction, and tech. Posts and comments are scraped from selected subreddits, cleaned, and analyzed for software/tool mentions. A Retrieval-Augmented Generation (RAG) chatbot will later be developed so users can query insights conversationally. The project also compares RAG performance with an LLM-only baseline.

**Project Timeline:**

Week 1 – Project definition and submit proposal (DONE)

Week 2 – Data collection (DONE)

Week 3 – Data cleaning, preprocessing, and initial keyword filtering (DONE)

Week 4 – Build software dictionary and extract mentions (DONE)

Week 5 – Sentiment analysis and identification of pain points (In Progress)

Week 6 – Develop and test RAG + LLM pipelines (Pending)

Week 7 – Evaluation of RAG vs LLM and refinement (Pending)

Week 8 – Final analysis, presentation preparation, and submission (Pending)

**Planned Work for the Week:**This week I planned to refine the dataset by building a software dictionary for each category (Law, Construction, Tech) and then apply this dictionary to extract mentions of tools from Reddit discussions. The goal was to identify relevant software and prepare a clean dataset for further analysis in Week 5.

**Progress for the Week:**

* Expanded the keyword dictionary to include domain-specific software/tools (e.g., Clio, Procore, Linux).
* Applied keyword detection to the dataset and flagged posts/comments containing tool mentions.
* Filtered the dataset to keep only relevant rows, reducing noise from unrelated conversations.
* Saved the cleaned dataset for downstream analysis.

**Roadblocks/Issues:**

* Some keywords overlap across categories (e.g., “cloud” can appear in law or tech contexts). I need to refine keyword lists further.
* Mentions of tools can be very short or vague, which may affect how clearly pain points are extracted.

**Plan for next Week:**

* Begin sentiment analysis on the cleaned dataset.
* Identify common pain points associated with frequently mentioned tools.
* Explore grouping mentions by subreddit to highlight category-specific insights (law firms vs. construction vs. tech)

**Resources for the Week:**  
No new external resources this week beyond the existing dataset and keyword dictionary.