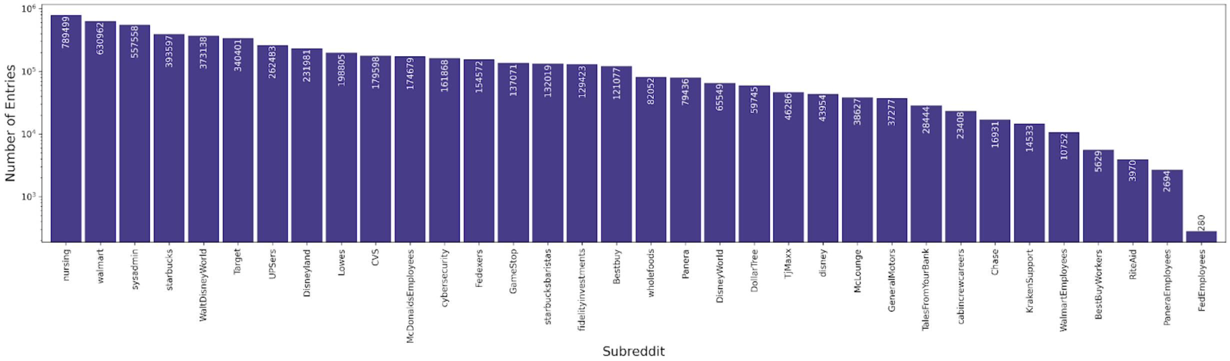
Project Summary, Stakeholders, and KPIs

Dataset: Aware dataset downloaded from reddit
Project description: We are given a dataset of 5+ Million entries covering 34 subreddits (the breakdown of the distribution of comments by subreddit can be seen in the figure below). The goal of this project is to construct a retrieval system designed to rank the most relevant content from this dataset for use in a RAG model. The priorities of this project are that the retrieval is fast (occurs sub-second), and that there is a methodology to gauge the performance of the retrieval.



Stakeholders:

- Aware: Jason Morgan (advisor)
- Team Members: Craig Franze, Himanshu Raj, Anil Tokmak, Mohammad Nooranidoost, Baian Liu, Peter Williams
- Future/current clients of Aware

KPI:

- Response time (needs to be subsecond)
- Scoring system (Recall, Precision, NDCG - Normalized discounted cumulative gain, other …)

Initial Questions:

1. What is the scope of prompts that we should expect the retrieval process to handle?
2. Should we anticipate prompts of a comparative nature (i.e. How does company X employee satisfaction compare to company Y?).
3. What is the ideal output structure for a retrieval model for this project?
    a. Just a list of the top n most relevant "chunks" as deemed by our model?
    b. Should we consider a reranking process for input into a llm?
4. How do we devise metrics for unranked data?
    a. Concept 1: create prompts/queries from chunks of the reddit data. Evaluate on the ability of the model to retrieve the original chunk from which the prompt was generated.
    b. Concept 2: Use LLM to give a binary value on whether a chunk contains information relevant to a given prompt. Run this over a large subset of chunks. Randomly sample the labeling and use human (our team) validation to assert its quality.