CS410 Final Project Progress Report
Hannah Benig - NETID: hhlim2

1) Which tasks have been completed?

- Understand how the NewsCatcherAPI works and set it up to gather articles and metadata based on a query (2 hours)
- Create a set of evaluation questions (1 hour)
- Set up LLM to receive context from news articles (1 hour)

2) Which tasks are pending?

- Evaluate LLMs from hugging face that we will subset with the news articles based on the answers to evaluation questions and ease of use (plan to try out Llama-2, Flan UL2, and roBERTa to see which works best) (5 hours)
- Create vector database implementing BM25 for ranking (4 hours) - now only using BM25 ranking, see note in challenges below.
- Fine-tune LLM parameters (temperature, top-p, top-k, etc.) to find which combination gives the best answer (3 hours)
- Create streamlit web app for frontend and package everything up for grading/evaluation (4 hours)

3) Are you facing any challenges?

- I ran into an issue with the Llama-2 model, it requires more compute resources than my computer could handle. I have been able to get around this issue using a smaller model but the answers are not as accurate. I am working to provide the model with better information in order to return correct answers.
- BM25 ranking is not commonly paired with a vector database. Instead, I am using BM25 to rank all documents returned by the NewsCatcherAPI.