CS410 Final Project Proposal - Hannah Benig

Requirements

1. **What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.**
    a. Captain: Hannah Benig (NetID: hhlim2)
2. **What topic have you chosen? Why is it a problem? How does it relate to the theme and to the class?**
    a. I have chosen the Intelligent Browsing Topic.
    b. Intelligent assistant answering questions from recent relevant news articles. Currently, if you are to ask a question in the google search engine, you may receive an answer to your question but will more likely receive a list of documents and then you need to read the documents to deem if they are relevant and find the answer to your question. In this project I plan to create a web application that uses an API to gather recent news articles and then index the gathered documents using a custom vector database that ranks documents using BM25 (week 2) so that users can query the application about recent news events. The output will be a direct answer to the question, the source(s) where the answer was found and the cosine similarity between the query and returned answer. This project will relate to the class in that I will be creating a collection language model (week 4) using recent news articles and will calculate the cosine similarity (week 1) between the query and returned answer.

3. **Briefly describe any datasets, algorithms or techniques you plan to use**
    a. I will use the NewsCatcherAPI to gather news articles (https://www.newscatcherapi.com/)
    b. I will create a vector database that uses BM25 to determine document ranking.
    c. I will subset a publicly available large language model from Huggingface with news articles to retrieve an answer to the user's query
    d. I will determine the best answer using cosine similarity
4. **How will you demonstrate that your approach will work as expected?**
    a. I will create a list of 5-10 sample questions with known answers and will determine that the approach has worked if 75% correctness has been achieved.
5. **Which programming language do you plan to use?**
    a. Python
6. **Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.**
    a. N=1, Total Workload ~20 hours
    b. Tasks to be completed and estimated time of completion
        i. Understand how the NewsCatcherAPI works and set it up to gather articles and metadata based on a query (2 hours)
        ii. Create vector database implementing BM25 for ranking (4 hours)
        iii. Create a set of evaluation questions (1 hour)
        iv. 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)
        v. Set up LLM to receive context from news articles (1 hour)
        vi. Fine-tune LLM parameters (temperature, top-p, top-k, etc.) to find which combination gives the best answer (3 hours)
        vii. Write cosine similarity function to calculate between query and answer (2 hours)

viii. Create streamlit web app for frontend and package everything up for grading/evaluation (4 hours)