Team Project Proposal

Project Title: Generative Search

1.Team Members:

- Team Captain: Mohammad Tamim (NetID: tamim2)
- Team Members:
  - Randi Weston (NetID: rweston2)
  - Timothy Cannella (NetID: tdc5)
  - Luke Freitag (NetID: lukegf2)

2. Project Description:

Our chosen topic for this project is "Generative Search". Our primary objective is to harness the capabilities of the Retrieval Augmented Generation (RAG) combined with Large Language Models (LLMs), to create a generative search system that sets new standards in information retrieval. The generative search system is a sophisticated tool capable of providing users with contextually accurate and informative responses to their queries. This system will be trained on the transcripts of the " CS 410 Text Information Systems" Coursera course.

Task: The core task of our project is to develop a search system that, when presented with a user's query related to the "CS 410 Text Information Systems" course material, generates accurate and informative responses.

Importance and Interest: The traditional search engines and information retrieval systems often present users with a list of search results that may require additional sifting and interpretation. This project seeks to address this limitation by creating a system that, when presented with a user's query, generates human-like responses that are contextually accurate and tailored to the query.

Thus, what makes our project particularly intriguing is its ability to go beyond mere retrieval and instead provide generative responses. This shift from retrieval to generation is of utmost importance as it ensures that the responses provided are not just relevant but are coherent and contextually accurate, thereby reducing the likelihood of the system generating false or misleading information. This addresses the challenge of hallucination often associated with LLMs.

Approach: Our planned approach includes:

- Collecting all the "CS 410 Text Information Systems" transcripts data from Coursera.
- Generating a high-quality question-and-answer (Q&A) dataset from "CS 410 Text Information Systems" Coursera transcripts.
- Developing the generative search system using Python in Google Colab.
- Evaluating the system's performance using cosine similarity to measure the proximity of generated responses to the actual content and manual reviews by team members.

Tools and Resources:

- Programming Language: Python
- Tools: Google Colab

- Datasets: CS 410 Text Information Systems transcripts from Coursera
- Model: RAG-based LLM

Expected Outcome: We anticipate that the system will provide accurate and contextually relevant responses to user queries related to the "CS 410 Text Information Systems" course.

Evaluation: We will create a set of 50 questions related to "CS 410 Text Information Systems" and evaluate the generative search response using:

- Quantitative Evaluation: We will employ cosine similarity as a quantitative metric to assess the proximity of the generated responses to the questions. A higher cosine similarity score indicates a more accurate and contextually relevant response. This quantitative evaluation will provide us with a measurable and objective assessment of the system's performance.

- Qualitative Evaluation: In addition to quantitative metrics, we recognize the importance of human judgment in evaluating the quality of the responses. Hence, we will conduct a manual review of the generated responses. Team members will assess responses for accuracy, and relevance with the original query. This qualitative evaluation ensures that the system not only provides technically sound answers but also maintains the context and meaning of the questions posed by users.

3. Programming language:

In the development of our generative search system, we will use Python programming language.

4. Justification of Workload

The justification for a workload of 97 hours is as follows:

- **Scraping "CS 410 Text Information Systems" Coursera transcripts**: This task is expected to take approximately 1 hour.
- **Performing data quality check and Data Cleaning on scraped transcripts**: This task is estimated to take about 8 hours.
- **Q&A dataset creation from scraped transcripts**: Compiling a high-quality dataset for training will require around 10 hours.
- **Design, Development, and testing of the generative search system**: This is a significant task and is estimated to take about 60 hours.
- **Results evaluation**: The process of quantitative and qualitative evaluation of search results requires roughly 10 hours.
- **Documentation**: Properly documenting the project, its design, and findings will need approximately 8 hours.

We are enthusiastic about this project and look forward to making significant contributions to the field of generative search. This endeavor not only challenges our technical skills but also aligns with the current trends in information retrieval, and natural language processing.

Thank you for considering our proposal.

Sincerely,
The Text Transformers