Automatic Search Result Evaluation using LLMs

Mark Ford 2551132F

December 15, 2023

1 Status Report

1.1 Proposal

1.1.1 Motivation

Building on the foundation of ”One-Shot Labeling for Automatic Relevance Estimation” by Sean MacAvaney and Luca Soldaini, this project seeks to address unresolved issues in using Large Language Models (LLMs) for automatic relevance estimation. The motivation is to enhance the efficiency and accuracy of offline search system evaluations, especially in scenarios involving long documents and both relevant and non-relevant documents as evaluation signals.

1.1.2 Aims

The project aims to:

1. Examine the application of LLMs in evaluating long documents and their effectiveness.
2. Investigate approaches for incorporating known relevant and non-relevant documents in the evaluation process.
3. Develop a practical framework for enhancing search result evaluations using LLMs.

1.2 Progress

- Replicated experiments from ”One-Shot Labeling for Automatic Relevance Estimation”
- Overcame challenges in setting up runtime environments and data importation.
- Encountered and resolved issues related to code execution in Google Colab.
- Implemented a caching mechanism for the duoT5 model.
- Began adapting code for the msmarco-document dataset.
- Explored data distributions for long document analysis.

1.3 Problems and Risks

1.3.1 Problems

- Technical challenges in code execution and environment setup.
- Difficulty in obtaining complete results for the msmarco-document dataset.
- Time constraints due to academic commitments.

1.3.2 Risks

- Potential for continued technical difficulties with the msmarco-document dataset. Mitigation: Focus on parallel tasks and other aspects of the project that do not rely solely on this dataset.
- Uncertainty in method selection due to project complexity. Mitigation: Prioritize in-depth exploration of a few selected methods rather than a superficial study of many.

2 Project Plan

2.1 Immediate Next Steps (Upcoming Weeks)

- **Technical Resolution:** Focus on resolving ongoing technical challenges and data acquisition issues with the msmarco-document dataset.
- **Initial Dissertation Drafting:**
  - Start drafting the *Introduction* chapter, outlining the motivation and scope of your project.
  - Begin compiling notes and sources for the *Background* chapter.

2.2 Following Month

- **Experimentation with Long Documents:**
  - Conduct experiments on long documents.
  - Refine methodologies based on initial findings and dataset characteristics.
- **Progress in Dissertation Writing:**
  - Develop the *Analysis/Requirements* chapter, detailing the problem definition and project requirements.
  - Start outlining the *Design* chapter, focusing on the abstract design of your solution.

2.3 Subsequent Month

- **Model Integration and Analysis:**
  - Analyze outcomes of experiments and explore other questions e.g. How do you handle multiple known relevant documents? If possible.

- Explore additional models or techniques as necessary.

- **Continued Dissertation Efforts:**
  - Expand on the *Implementation* chapter, documenting your technical achievements and development process.
  - Begin formulating the *Evaluation* strategy and criteria.

2.4 Final Month

- **Finalizing and Testing:**
  - Finalize research findings and complete any remaining experiments.
  - Prepare for a comprehensive evaluation of your solution.
- **Dissertation Completion:**
  - Complete the *Evaluation* chapter with results and analysis.
  - Write the *Conclusion* chapter, summarizing the project and suggesting future work.
  - Revise and polish all chapters, ensuring coherence and completeness.
  - Prepare appendices and any additional documentation required.
  - Submit the first complete draft to your supervisor for feedback.

2.5 Post-Completion

- **Revisions and Final Submission:**
  - Incorporate feedback from your supervisor and make necessary revisions.
  - Complete final proofreading and formatting adjustments.
  - Submit the final version of your dissertation.

2.6 Ethics and Data Considerations

- Since the project does not involve human subjects or sensitive data, continue to ensure that no ethical approval is required. Maintain focus on publicly available documents and model-generated evaluations.