Northeastern University, Khoury College of Computer Science

CS 6120 Natural Language Processing - Final Project

Due: April 24, 2025(100 points)

NAME(S)

PROJECT GIT REPOSITORY

The project is 25% of your grade, and is broken down as follows. The report comprises 33% of your project grade and the technical artifacts, demonstration, and real-time capabilities comprise 66%. More details about the presentation breakdown can be found on the website.

1 Objectives and Introduction

1.1 RAG LLM Projects

Introduce us to the capability that you have built, the domain in which you've selected your data, and the objective that you're attempting to achieve with your LLM.

1.2 Publication Submissions

You may use whatever template is appropriate for the conference that you are considering.

Almost universally, conference proceedings contain an introduction section. Here, you will typically introduce your problem, describe why it's important and interesting, and why it's worth building a solution to. Incorporate elements of your proposal, adding both your objective and motivation. Ensure that you mention the impact the work will have and most importantly, your particular contribution. Other perspectives to touch upon in the introduction include the challenges that state of the art faces. Finally, outline the remainder of your document.

The `Introduction` section is typically the entirety of the front matter (and oftentimes bleeds into the second page of your publication.) It is oftentimes the most a reader will read, and so it is important to be engaging and to the point.

2 Background and Related Work

2.1 LLM RAG Projects

Provide detailed perspectives of the data and its origin. What are the attributes of this data? Why is it interesting? Are there structured portions of the data that helps you with your chunking? Most importantly, provide a link and/or instructions on how anyone can obtain this data, since it is unlikely that you will be able to store it in your repositories.

2.2 Publication Submissions

This section should describe how this problem arises, who is affected, and any literature that you plan to review or survey. These references should relate to both the applications of the work as well as the technical approaches to be taken. Emphasize why they are appropriate and relevant.

At times this section is titled `Related Work` or incorporates cited literature. Incorporate these references liberally and group them to show the breadth of your survey.

3 Approach and Implementation

3.1 RAG LLM Projects

At the end of this section, include the link to your repository's README.md. It should be evident on how to replicate the LLM capability. Things that will help you is:

- **Dockerization**: Containerization of your application will allow TA's and Instructor to assess is reproducibility.
- **Scripting**: Your Dockerfile can have a script that will automatically run every time you run `docker run`.
- **Automated Download**: Inside your script, you should be able to automate the download and processing of the corpus. Provided you have appropriate attribution, you can either process the data in situ or use your own server space (or even Google Drive if it is less than 25GB) to host your processed dataset.

In this section, include the general approach that you have selected, the design decisions, and the considerations that have led to your implementation. Include any novel algorithm contributions that you required in training and deploying the system.

3.2 Publication Submissions

This section should provide the meat of your work. There should be a *narrative* to your approach rather than a collection of surveyed results that you may have implemented. The project should be a cohesive description of the approaches and algorithms and how they relate to each other.

As well, ensure that it is clear what your contribution is. That is, what is the *novel contribution* that adds upon existing work?

4 Data and Data Analysis

4.1 Data Source(s)

- Please cite the dataset that you have used in either LLM RAG system or paper submission, including proper citation of authorship and website URL. Regarding the dataset scale, there should be enough rows / samples for the mining to be interesting and the conclusions to be statistically significant. This number varies with number of features and labels, but more than 100,000 rows is generally sufficient in most cases (more is better). It must not be able to be intuited by hand, at least easily.

4.2 Data Analysis and Exploration

- Please analyze your data before doing any modeling. Understand the distributions of your data. Determine feature correlation with the label (if you have labels) and each other.

5 Results and Evaluation

Clearly define the evaluation metrics you're using and why. Properly evaluate and draw conclusions. Determine the confidence you have in the outcomes of your processing and modeling. Ensure that you have subjective examples in this section as well.

6 Conclusions

Discuss what you've learned. How do you think it can be applied? If you had more time, what are the future directions? If there's literature on future directions, include them.

7 Submission Guidelines

Upload your PDF file, GCP (or other) Endpoint Link, and Github URL to Gradescope as a *single* submission with all members of the team. On the top of the document, include the teammates and project repository.