LLM Fine-tuning : Impression Generation

Objective

Demonstrate proficiency in LLM, and NLP techniques, including model fine-tuning, text analysis, and data visualization, using the given dataset.

Assignment Details

1. Model Fine-tuning

- Choose between the gemma-7b-it or gemma-2b-it model based on your available hardware capabilities.
- Fine-tune the selected model to generate Impressions based on the Report Name,History,Observation given in the dataset.
- Use 300 samples from the dataset for training and reserve 30 samples for evaluation.

2. Model Evaluation

- Generate impressions using your fine-tuned model on the 30 evaluation samples.
- Compute and report the following metrics:
  a. Perplexity
  b. ROUGE score

3. Text Analysis

Perform the following analysis on the entire dataset of 330 reports:

- Remove stop words from the text
- Apply stemming and lemmatization to the remaining words
- Convert the processed text into embeddings
- Identify the top 100 pairs of words based on embedding similarity

4. Visualization

- Create a visualization of the top 100 word pairs identified in the text analysis step.
- Bonus: Develop an interactive visualization for exploring these word pairs.

Deliverables

1. A GitHub repository containing:
    - All source code used for model fine-tuning, evaluation, text analysis, and visualization
    - Documentation explaining your approach, methodologies, and any assumptions made
    - Results of the model evaluation (perplexity and ROUGE scores)
    - Visualization(s) of the top 100 word pairs
    - (If completed) Interactive visualization code or link
2. A brief report summarizing your findings, challenges encountered, and potential areas for further improvement

Submission Guidelines

- Ensure your GitHub repository is public and contains all required deliverables.
- Submit the link to your GitHub repository using the provided submission form.
- Deadline: Submit your work no later than 1:00 PM on the specified due date.

Additional Notes

- You may use any Python libraries or frameworks that you find appropriate for this task.

- Clearly document any third-party code or pre-trained models used in your solution.
- If you make any assumptions about the data or task, clearly state them in your documentation.

Good luck, and we look forward to reviewing your innovative solutions!