Project Guidelines

Issue Date: 2nd July 2024

Your class project is an opportunity for you to explore an interesting problem relevant to large language models. Each one of you will work in a group of two students on a single project. One or three student groups are exceptional and need instructor permission. The instructor and teaching assistant will consult with you on your ideas (if necessary), but of course, the final responsibility to define and execute an interesting piece of work is your responsibility. Your project will be worth 30% of your final class grade. Each group chose the project from the provided list of topics or proposed their own idea. It is possible, that more than one group can select the same dataset but need to apply different techniques. Please provide the details of your project in the following link.

https://docs.google.com/spreadsheets/d/193VEmSMfBUv4JsnkxmNRsthW2zmggqZ05ri5IP3X-A8/edit?usp=sharing

Expectation: Each project should be developed in python and rigorous analysis is required.

Deliverables

1. **Preliminary Submission**

   a) **Proposal: 1 page document [Points: 10]**

   - Title with group member names
   - Project idea (should be understandable to anyone in the course)
   - Method/Technique
   - Dataset explanation and accessible link
   - Timeline with each member individual contribution.
   - References

   Template: used which one suits you well (i.e., MS Word, Latex, Overleaf) from the following link: https://www.ieee.org/conferences/publishing/templates.html

   Submission: Upload *.pdf* proposal in Moodle (Only the team lead will submit it on behalf of the team – Don't forget to mention the name of the group members inside the report to grade each of them)
   Deadline: 5th July 2024

2. **Project Check Point [Points: 20]**
   a. This deliverable is to check the progress made so far in the project. The requirement is at least 50% portion of work to be done.
   Submission: Github repo link
   Deadline: 14th July 2024

3. **Final Submission**

   a. **Project Report [Points: 30]**

   (Template: https://www.ieee.org/conferences/publishing/templates.html)

   i. Introduction
   ii. Related Work
   iii. Methodology
   iv. GitHub link: (Double check the link it should work properly)
   v. Experiments and Evaluation
   vi. Analysis and Observations
   vii. Conclusion
   viii. Reference

   Submission: Upload *.pdf* report in Moodle (Minimum page limit: 4 Pages) – Only team lead will upload the report on behalf of the team.

   b. **GitHub Maintenance [Points: 10]**

   i. .md file explaining the project details such as title and one paragraph of about the project. (Strictly follow this guideline)
   ii. Good coding practice should be followed

   Deadline: 22nd July 2024

   c. **Final Presentation: [Points: 10]**

   i. Duration: 8 Minute (Effective time management is crucial for this task. Failure to complete within 8 minutes will result in a penalty of 5 points)
   ii. Follow the shared template in Moodle
   Deadline: During the class time – **23rd July 2024**
   (Schedule will be shared).

   d. **Project Demo [Points: 20] – Deadline: 23rd July 2024**
   (Schedule will be shared accordingly)

NOTE: If your code is not working during the project demo then your whole submission will be evaluated in 50% marks of the "Final Submission".

- **Policy**
  - **No late submission is allowed.**
  - Your code should be properly commented
  - If (plagiarized code == True || code != working) then "0"

List of Topics

a. LLM-Based Chatbots for Mining Software Repositories
   Source: https://das.encs.concordia.ca/pdf/abedu2024llm.pdf

b. Chatbots
   Sources: https://realpython.com/build-llm-rag-chatbot-with-langchain/
   https://medium.com/@danushidk507/creating-a-simple-chatbot-with-open-source-llms-using-python-and-hugging-face-01a9f5a7ebdf

c. Document Question Answering
   Source: https://huggingface.co/tasks/document-question-answering

d. Sentence Similarity
   Source: https://huggingface.co/tasks/sentence-similarity

e. Text Classification
   Source: You can find lot of resources over the internet.

f. LLM pruning for fine-tuning
   Source: https://web.stanford.edu/class/cs224n/final-reports/256911740.pdf

g. Text Summarization
   Source: https://github.com/shivaniNK8/News-Article-Text-Summarizer-Transformer

********** Good Luck **********