CS 613: NLP

Assignment 2: Tokenizer & Model Training

Total marks: **100 Pts.**	**Submission deadline: 23:59:59 Hrs, November 14, 2024 (Extended)**

Assignment Instructions

1. Regarding the late submission, we will be following the penalty as per the table:

Late Submission	**Penalty (Out of 100)**
Till 1-hour past the deadline	5 points
1 to 12 hours past the deadline	10 points
12 to 24 hours past the deadline	20 points
24 to 36 hours past the deadline	40 points
36+ hours past the deadline	100 points

2. We will follow the zero plagiarism policy, and any act of plagiarism will result in a zero score for the assignment.
3. Please cite and mention others' work and give credit wherever possible.
4. If you seek help and discuss it with the stakeholders or individuals, please ask their permission to mention it in the report/submission.
5. Compute requirement: Use Colab and write the answers in the Colab itself.

Problem Statement (100 Points)

Task 1: Tokenizer Training

1. Train 5 Tokenizers on five samples from the dataset you had scraped from the earlier assignment. **(25 Pts)**
2. Calculate the fertility score of all the five Tokenizers that you have trained. **(20 Pts)**
3. Show the matrix with the fertility score and dataset size. **(05 Pts)**

Task 2: Model Training

1. Choose any of the predefined model architectures & adjust it in such a way that its total parameters are less than 100M. **(25 Pts)**
2. Tokenize your dataset using the best tokenizer you trained in Task 1. **(05 Pts)**
3. Train your model using the tokenized dataset. Note down the perplexity of your model for every 0.1 epoch. **(10 Pts)**
4. Show the matrix with perplexity for each epoch and test the model's output for 10 prompts. **(10 Pts)**

Submission

1. Submit your code (GitHub) or colab notebook with proper comments to this link.
    a. Ensure the individual contribution is appropriately added (OTHERWISE PENALTY OF 10 MARKS).

Expectations from the team:

1. Properly divide the team into sub-groups and distribute your tasks equally.
2. Negative marks for documentation and justifications!
3. Write the contributions or tasks completed by each team member. Scores might be different among team members if the tasks are not equally distributed.

References

1. https://github.com/AamodThakur/NLP_Pre_Training

TAs to Contact

1. Himanshu Beniwal (himanshubeniwal@iitgn.ac.in)
2. Indrayudh Mandal (24210041@iitgn.ac.in)
3. Mithlesh Singla (24210063@iitgn.ac.in)
4. Alay Patel (alay.patel@iitgn.ac.in)
5. Aamod Thakur (aamod.thakur@iitgn.ac.in)

FAQs

1. *We will add clarifications to doubts here. Please check periodically, as someone might have already asked about the doubt, which will be appended here.*