Take-Home Assignment: Tokenization and Byte Pair Encoding

Deadline: 7 March

1 Objective

This assignment aims to explore the Byte Pair Encoding (BPE) compression ratio across multiple languages and compare it with the compression efficiency of GPT tokenization methods (GPT-2 and GPT-4). Additionally, students will analyze how file size impacts BPE compression ratios for English text.

2 Tasks

2.1 Task 1: Compute BPE Compression Ratios

- Start with Shakespeare Data in English, German, Spanish, and French. Dataset for all languages has been provided to you.
- Apply Byte Pair Encoding (BPE) tokenization on each language. We have seen the code for this in our lectures.
- Note that the final vocabulary size you can consider = Original vocabulary size + 200 extra tokens.
- Compute the compression ratio for each language.

2.2 Task 2: Bar Plot Comparison of BPE Compression Ratios

- Create a bar plot visualizing the BPE compression ratio across the four languages.
- Analyze and describe any observed trends.

2.3 Task 3: Comparison with GPT Tokenization Methods

- Compute the compression ratios using the Tiktoken library.
- Obtain compression ratios using Tiktoken library for GPT-2, GPT-3.5, and GPT-4 for all languages considered.
- Compare the compression efficiencies of BPE, GPT-2, and GPT-4 using bar plots for all languages considered.

2.4 Task 4: Explore Effect of Final Vocabulary Size on Compression Ratio

- Modify the final vocabulary size to test different limits: 200, 500, and 800 extra tokens.
- Compute the compression ratio for each setting.
- Analyze how the vocabulary size influences the compression ratio across different languages.
- Visualize the results using plots.

2.5 Task 5: Effect of File Size on BPE Compression Ratios

- Select an English text sample.
- Create text files with decreasing sizes using the scaling factors: 10, 8, 6.
- Note that the final vocabulary size you can consider = Original vocabulary size + 5% of the total text size.
- Compute BPE compression ratios for each file size.
- Plot a graph showing the effect of file size on BPE compression ratio.
- Provide an analysis of how file size impacts compression.

3 Submission Guidelines

1. **Detailed Report** (PDF format):
   - Introduction to BPE tokenization and compression ratio.
   - Explanation of methodology and approach used.
   - Results, plots, and analysis for each task.
   - Share your final code file.
   - Observations and conclusions.
   - References (if any external sources are used).
2. **Deadline:** 7 March. Late submissions will not be considered.

4 Team Size

You can have a maximum team size of 2 students per team. If you are not forming a team, you can submit the assignment individually.

5 Evaluation Criteria

- **Correctness of Implementation** (40%)
- **Clarity of Explanation & Analysis** (30%)
- **Quality of Plots & Visualizations** (20%)
- **Presentation & Report Structure** (10%)

For any clarifications, please reach out before the deadline. Good luck!