**Evaluation Metrics Report**

This report summarizes the evaluation metrics for the custom BPE tokenizer.

**Final Vocabulary Size:**

The final vocabulary size for the tokenizer is of 30,000 tokens.

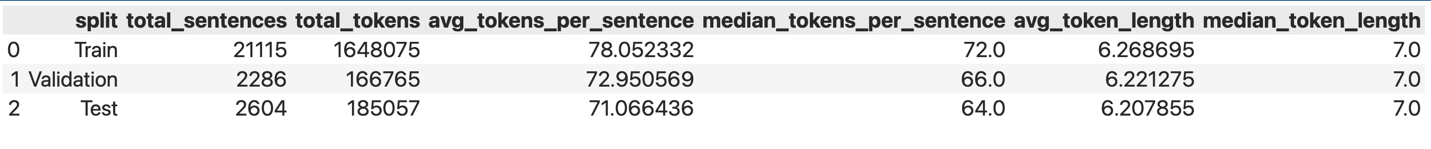
***Code to train the tokenizer and print the vocabulary size***

**A screenshot of a computer program

AI-generated content may be incorrect.**

**Average Tokens per Sentence Tokenization Consistency**

The table below shows an average of 78, 72, and 71 tokens per sentence in the training, validation, and testing sets, respectively. These numbers suggest slightly shorter sentences in the validation and test sets compared to the training set. Each dataset had an average token length of about six characters and a median value of seven tokens, indicating a stable average token length across all data splits.

****

***Code to generate the dataframe:***

A screenshot of a computer code

AI-generated content may be incorrect.

**Compression ratio**

The compression rate in terms of characters per tokens and bytes per tokens was found to be 6.21 and 6.22 respectively.

***Code to calculate the compression ratio***

**A computer screen shot of a computer code

AI-generated content may be incorrect.**