Report

Unfortunately, I didn't manage to complete the working version of the code. During the task, I encountered several issues that I wasn't able to resolve in time, but I plan to finish the project in the near future.

I have selected a suitable dataset for solving the task of identifying mountain names. However, I didn't perform a detailed analysis of this dataset, so I cannot confirm its quality or whether it needs preprocessing or improvements before use. It would be helpful to conduct basic data analysis first, check the class distribution, the number of positive examples, and identify potential labeling issues.

It's important to analyze the class distribution in the training, validation, and test datasets before training the model. This will help better understand the class balance and determine whether class weighting needs to be applied to correct any imbalance. For example, I should check if there are enough positive examples (mountain names) to ensure effective model training.

When preparing the data, it's worth experimenting with different tokenization methods such as WordPiece or Byte-Pair Encoding (BPE) to choose the best one for this specific task. Different tokenizers can handle text differently and influence the results, especially when identifying specific geographic names.

In addition to using `dslim/bert-large-NER`, I should try other models for named entity recognition, such as RoBERTa, ALBERT, or T5. Each of these architectures may have its own advantages depending on the characteristics of the text and the task of recognizing mountain names.