NER_POTENTIAL_IMPROVEMENTS

1. Data Augmentation

- **Idea:**Increase the diversity of the dataset by applying data augmentation techniques, such as:
  - Generating synthetic sentences with mountain names in different contexts.
  - Using paraphrasing techniques to create variations of existing sentences.
- **Goal:** Improve the model's generalization ability by exposing it to a broader variety of sentence structures and contexts.

2. Increase Dataset Size

- **Idea:** Collect more labeled data by including additional real-world examples from different sources (e.g., travel articles, mountaineering guides, geographic databases).
- **Goal:** A larger dataset can help the model learn better patterns and improve its overall accuracy.

3. Use a More Advanced Pretrained Model

- **Idea:** Maybe instead of using BERT, we could try fine-tuning a more advanced pretrained model such as RoBERTa, XLNet, or DeBERTa.
- **Goal:** These models often have better contextual understanding and can improve the performance of the NER task.

4. Post-Processing with Rules

- **Idea:**We could add some rules to help to identify tokens into coherent multi-word mountain names.
  - Example: If the model predicts "Mount" as B-MOUNTAIN and "Everest" as I-MOUNTAIN, merge them into a single entity "Mount Everest".
- **Goal:** Reduce false negatives and improve F1-Score for multi-word names.

Conclusion

By applying these improvements, the model's accuracy and robustness can be increased and we could get better results on inference part.