Xử lý dữ liệu văn bản

Hướng dẫn này hiển thị các phương pháp cụ thể để xử lý tập dữ liệu văn bản. Tìm hiểu cách:

- Tokenize a dataset with map().
- Căn chỉnh nhãn tập dữ liệu với id nhãn cho tập dữ liệu NLI.

Để biết hướng dẫn về cách xử lý bất kỳ loại tập dữ liệu nào, hãy xem hướng dẫn quy trình chung.

Bản đồ

The map() function supports processing batches of examples at once which speeds up
token hóa.

Tải mã thông báo từ Transformers:

```
>>> from transformers import AutoTokenizer

>>> tokenizer = AutoTokenizer.from_pretrained("bert-base-cased")
```

Set the batched parameter to True in the map() function to apply the tokenizer to batches of
ví dụ:

```
>>> dataset = dataset.map(lambda examples: tokenizer(examples["text"]), batched=True)
>>> dataset[0]
{'text': 'the rock is destined to be the 21st century\'s new " conan " and that he\'s going to mak
  'nhãn': 1,
  'input_ids': [101, 1996, 2600, 2003, 16036, 2000, 2022, 1996, 7398, 2301, 1005, 1055, 2047, 1000,
  'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
```

The map() function converts the returned values to a PyArrow-supported format. But explicitly
trả về các tensor dưới dạng mảng NumPy nhanh hơn vì đây là PyArrow được hỗ trợ nguyên bản
format. Set return_tensors="np" when you tokenize your text:

```
>>> dataset = dataset.map(lambda examples: tokenizer(examples["text"], return_tensors="np"), batch
```

Căn chỉnh

The align_labels_with_mapping() function aligns a dataset label id with the label name. Not all
Các mô hình Transformers tuân theo ánh xạ nhãn quy định của tập dữ liệu gốc, đặc biệt
cho bộ dữ liệu NLI. Ví dụ: tập dữ liệu MNLI sử dụng ánh xạ nhãn sau:

```
>>> label2id = {"entailment": 0, "neutral": 1, "contradiction": 2}
```

Để căn chỉnh ánh xạ nhãn tập dữ liệu với ánh xạ được mô hình sử dụng, hãy tạo một từ điển gồm
tên nhãn và id để căn chỉnh:

```
>>> label2id = {"contradiction": 0, "neutral": 1, "entailment": 2}
```

Pass the dictionary of the label mappings to the align_labels_with_mapping() function, and the
cột để căn chỉnh trên:

```
>>> from datasets import load_dataset

>>> mnli = load_dataset("nyu-mll/glue", "mnli", split="train")
>>> mnli_aligned = mnli.align_labels_with_mapping(label2id, "label")
```

Bạn cũng có thể sử dụng chức năng này để gán ánh xạ nhãn tùy chỉnh cho id.