Tải dữ liệu văn bản

Hướng dẫn này chỉ cho bạn cách tải tập dữ liệu văn bản. Để tìm hiểu cách tải bất kỳ loại tập dữ liệu nào, hãy
hãy xem hướng dẫn tải chung.

Tệp văn bản là một trong những loại tệp phổ biến nhất để lưu trữ tập dữ liệu. Theo mặc định, Bộ dữ liệu
lấy mẫu từng dòng tệp văn bản để xây dựng tập dữ liệu.

```
>>> from datasets import load_dataset
>>> dataset = load_dataset("text", data_files={"train": ["my_text_1.txt", "my_text_2.txt"], "test"

# Tải từ một thư mục
>>> dataset = load_dataset("text", data_dir="path/to/text/dataset")
```

Để lấy mẫu tệp văn bản theo đoạn văn hoặc thậm chí toàn bộ tài liệu, hãy sử dụng tham số sample_by:

```
# Mẫu theo đoạn văn
>>> dataset = load_dataset("text", data_files={"train": "my_train_file.txt", "test": "my_test_file

# Lấy mẫu theo tài liệu
>>> dataset = load_dataset("text", data_files={"train": "my_train_file.txt", "test": "my_test_file
```

Bạn cũng có thể sử dụng mẫu grep để tải các tệp cụ thể:

```
>>> from datasets import load_dataset
>>> c4_subset = load_dataset("allenai/c4", data_files="en/c4-train.0000*-of-01024.json.gz")
```

Để tải các tệp văn bản từ xa qua HTTP, thay vào đó hãy chuyển các URL:

```
>>> dataset = load_dataset("text", data_files="https://huggingface.co/datasets/hf-internal-testing
```

Để tải dữ liệu XML, bạn có thể sử dụng trình tải "xml", tương đương với "văn bản" với
sample_by="document":

```
>>> from datasets import load_dataset
>>> dataset = load_dataset("xml", data_files={"train": ["my_xml_1.xml", "my_xml_2.xml"], "test": "

# Tải từ một thư mục
>>> dataset = load_dataset("xml", data_dir="path/to/xml/dataset")
```