Cấu trúc kho lưu trữ của bạn

Để lưu trữ và chia sẻ tập dữ liệu của bạn, hãy tạo kho lưu trữ tập dữ liệu trên Hugging Face Hub và tải lên các tập tin dữ liệu của bạn.

Hướng dẫn này sẽ chỉ cho bạn cách cấu trúc kho lưu trữ tập dữ liệu khi bạn tải nó lên.
A dataset with a supported structure and file format (.txt , .csv , .parquet , .jsonl , .mp3 , .jpg , .zip etc.) are loaded automatically with load_dataset(), and it'll have a dataset viewer trên trang dữ liệu của nó trên Hub.

Trường hợp sử dụng chính

The simplest dataset structure has two files: train.csv and test.csv (this works with any supported file format).

Kho lưu trữ của bạn cũng sẽ chứa tệp README.md, thẻ tập dữ liệu được hiển thị trên tập dữ liệu của bạn trang.

```
my_dataset_repository/
 README.md
 train.csv
 test.csv
```

In this simple case, you'll get a dataset with two splits: train (containing examples from train.csv) and test (containing examples from test.csv).

Xác định phần tách và tập hợp con của bạn trong YAML

Chia tách

Nếu bạn có nhiều tệp và muốn xác định tệp nào sẽ được chia thành phần nào, bạn có thể sử dụng Trường cấu hình YAML ở đầu README.md của bạn.

Ví dụ: đưa ra một kho lưu trữ như thế này:

```
my_dataset_repository/
 README.md
 dữ liệu.csv
 holdout.csv
```

Bạn có thể xác định phần phân chia của mình bằng cách thêm trường cấu hình trong khối YAML ở đầu README.md:

```
None
cấu hình:
- config_name: mặc định
    dữ liệu_files:
    - chia: tàu
        đường dẫn: "data.csv"
    - chia: kiểm tra
        đường dẫn: "holdout.csv"
None
```

Bạn có thể chọn nhiều tệp cho mỗi lần chia bằng danh sách đường dẫn:

```
my_dataset_repository/
 README.md
 dữ liệu/
  abc.csv
  def.csv
 kiên trì/
        ghi.csv
```

```
None
cấu hình:
- config_name: mặc định
  dữ liệu_files:
  - chia: tàu
    con đường:
    - "dữ liệu/abc.csv"
    - "dữ liệu/def.csv"
  - chia: kiểm tra
    đường dẫn: "holdout/ghi.csv"
None
```

Hoặc bạn có thể sử dụng các mẫu toàn cầu để tự động liệt kê tất cả các tệp bạn cần:

```
None
cấu hình:
- config_name: mặc định
  dữ liệu_files:
  - chia: tàu
    đường dẫn: "data/*.csv"
  - chia: kiểm tra
    đường dẫn: "holdout/*.csv"
None
```

[!WARNING]
Lưu ý rằng trường config_name là bắt buộc ngay cả khi bạn có một cấu hình duy nhất.

Cấu hình

Tập dữ liệu của bạn có thể có một số tập hợp con dữ liệu mà bạn muốn tải riêng. TRONG
trong trường hợp đó, bạn có thể xác định danh sách cấu hình bên trong trường configs trong YAML:

```
my_dataset_repository/
 README.md
 main_data.csv
 bổ sung_data.csv
```

```
None
cấu hình:
- config_name: dữ liệu chính
    data_files: "main_data.csv"
- config_name: dữ liệu bổ sung
    data_files: "bổ sung_data.csv"
None
```

Mỗi cấu hình được hiển thị riêng biệt trên Hugging Face Hub và có thể được tải bằng
chuyển tên của nó làm tham số thứ hai:

```
từ tập dữ liệu nhập Load_dataset

main_data = load_dataset("my_dataset_repository", "main_data")
additional_data = load_dataset("my_dataset_repository", "additional_data")
```

Thông số trình tạo

Không chỉ data_files mà các tham số dành riêng cho trình tạo khác cũng có thể được chuyển qua YAML, cho p
để linh hoạt hơn về cách tải dữ liệu trong khi không yêu cầu bất kỳ mã tùy chỉnh nào. Ví dụ,
xác định dấu phân cách nào sẽ sử dụng trong cấu hình nào để tải tệp csv của bạn:

```
None
cấu hình:
- config_name: tab
    data_files: "main_data.csv"
    tháng chín: "\t"
- config_name: dấu phẩy
    data_files: "bổ sung_data.csv"
    tháng chín: ","
None
```

Tham khảo tài liệu của các nhà xây dựng cụ thể để xem họ có những thông số cấu hình nào.

```
[!TIP]
Bạn có thể đặt cấu hình mặc định bằng cách sử dụng default: true , ví dụ: bạn có thể chạy
```

main_data = load_dataset("my_dataset_repository") if you set

- config_name: dữ liệu chính
  data_files: "main_data.csv"
  mặc định: đúng

Tự động phát hiện sự phân chia

Nếu không cung cấp YAML, Bộ dữ liệu sẽ tìm kiếm các mẫu nhất định trong kho lưu trữ tập dữ liệu để tự động suy ra sự phân chia tập dữ liệu.
Có một thứ tự cho các mẫu, bắt đầu bằng định dạng phân chia tên tệp tùy chỉnh để xử lý tất cả các tập tin dưới dạng một phần tách nếu không tìm thấy mẫu nào.

Tên thư mục

Các tệp dữ liệu của bạn cũng có thể được đặt vào các thư mục khác nhau có tên train , test và xác thực trong đó mỗi thư mục chứa các tệp dữ liệu cho phần phân chia đó:

```
my_dataset_repository/
 README.md
 dữ liệu/
        luyện tập/
         bees.csv
        kiểm tra/
         thêm_bees.csv
        xác nhận/
              even_more_bees.csv
```

Tách tên tệp

Nếu bạn không có bất kỳ phần chia tách phi truyền thống nào thì bạn có thể đặt tên phần tách ở bất kỳ đâu t
tập tin dữ liệu và nó được tự động suy ra. Quy tắc duy nhất là tên phân tách phải được phân cách bằng các ký tự không phải từ, chẳng hạn như test-file.csv thay vì testfile.csv . Được hỗ trợ dấu phân cách bao gồm dấu gạch dưới, dấu gạch ngang, dấu cách, dấu chấm và số.

Ví dụ: tất cả các tên tệp sau đều được chấp nhận:

- chia tàu: train.csv , my_train_file.csv , train1.csv
- phân chia xác thực: validation.csv , my_validation_file.csv , validation1.csv
- phân tách kiểm tra: test.csv , my_test_file.csv , test1.csv

Đây là một ví dụ trong đó tất cả các tệp được đặt vào một thư mục có tên data :

```
my_dataset_repository/
 README.md
 dữ liệu/
        train.csv
        test.csv
        xác thực.csv
```

Tách tên tệp tùy chỉnh

Nếu phần tách tập dữ liệu của bạn có tên tùy chỉnh không phải là train , test hoặc validation , thì bạn can name your data files like data/<split_name>-xxxxx-of-xxxxx.csv .

Dưới đây là một ví dụ với ba phần tách, train, test và ngẫu nhiên:

```
my_dataset_repository/
 README.md
 dữ liệu/
        train-00000-of-00003.csv
        train-00001-of-00003.csv
        train-00002-of-00003.csv
        test-00000-of-00001.csv
        ngẫu nhiên-00000-of-00003.csv
        ngẫu nhiên-00001-of-00003.csv
        ngẫu nhiên-00002-of-00003.csv
```

Chia đơn

Khi Bộ dữ liệu không thể tìm thấy bất kỳ mẫu nào ở trên thì nó sẽ coi tất cả các tệp là một tàu chia cắt. Nếu phần tách tập dữ liệu của bạn không tải như mong đợi, có thể là do cài đặt không chính xác mẫu.

Tách từ khóa tên

Có một số cách đặt tên cho phần chia tách. Phân chia xác thực đôi khi được gọi là "dev" và kiểm tra sự phân chia có thể được gọi là "eval".
Các tên phân chia khác này cũng được hỗ trợ và các từ khóa sau là tương đương:

- huấn luyện, huấn luyện
- xác thực, hợp lệ, val, dev
- kiểm tra, kiểm tra, đánh giá, đánh giá

Cấu trúc bên dưới là một kho lưu trữ hợp lệ:

```
my_dataset_repository/
 README.md
 dữ liệu/
        đào tạo.csv
        eval.csv
        hợp lệ.csv
```

Nhiều tập tin mỗi lần chia

Nếu một trong các phần tách của bạn bao gồm nhiều tệp, Bộ dữ liệu vẫn có thể suy ra liệu đó có phải là đoà xác thực và kiểm tra tách khỏi tên tệp.
Ví dụ: nếu phần chia tách đào tạo và kiểm tra của bạn trải rộng trên nhiều tệp:

```
my_dataset_repository/
 README.md
 train_0.csv
 train_1.csv
 train_2.csv
 train_3.csv
 test_0.csv
 test_1.csv
```

Make sure all the files of your train set have train in their names (same for test and validation).
Even if you add a prefix or suffix to train in the file name (like my_train_file_00001.csv for

example),
Bộ dữ liệu vẫn có thể suy ra cách phân chia phù hợp.

Để thuận tiện, bạn cũng có thể đặt các tệp dữ liệu của mình vào các thư mục khác nhau. Trong trường hợp này, tên phân chia được suy ra từ tên thư mục.

```
my_dataset_repository/
 README.md
 dữ liệu/
        luyện tập/
         shard_0.csv
         shard_1.csv
         shard_2.csv
         shard_3.csv
        kiểm tra/
              shard_0.csv
              shard_1.csv
```