Tạo tập dữ liệu tài liệu

Hướng dẫn này sẽ chỉ cho bạn cách tạo tập dữ liệu tài liệu bằng PdfFolder và một số siêu dữ liệu. Đây là giải pháp không cần mã để tạo nhanh tập dữ liệu tài liệu với một số nghìn pdf.

[!TIP]
Bạn có thể kiểm soát quyền truy cập vào tập dữ liệu của mình bằng cách yêu cầu người dùng chia sẻ liên
thông tin đầu tiên. Hãy xem hướng dẫn về bộ dữ liệu Gated để biết thêm thông tin về cách kích hoạt tính năng này trên Hub.

Thư mục Pdf

PdfFolder là trình tạo tập dữ liệu được thiết kế để tải nhanh tập dữ liệu tài liệu với một số nghìn pdf mà không yêu cầu bạn phải viết bất kỳ mã nào.

[!TIP]
Hãy xem hệ thống phân cấp Mẫu phân tách để tìm hiểu thêm về cách tạo PdfFolder phân chia tập dữ liệu dựa trên cấu trúc kho lưu trữ tập dữ liệu của bạn.

PdfFolder tự động suy ra nhãn lớp của tập dữ liệu của bạn dựa trên tên thư mục.
Lưu trữ tập dữ liệu của bạn trong cấu trúc thư mục như:

```
thư mục/train/sơ yếu lý lịch/0001.pdf
thư mục/train/sơ yếu lý lịch/0002.pdf
thư mục/train/sơ yếu lý lịch/0003.pdf

thư mục/tàu/hóa đơn/0001.pdf
folder/train/invoice/0002.pdf
folder/train/invoice/0003.pdf
```

If the dataset follows the PdfFolder structure, then you can load it directly with load_dataset():

```
>>> from datasets import load_dataset

>>> dataset = load_dataset("path/to/folder")
```

This is equivalent to passing pdffolder manually in load_dataset() and the directory in dữ liệu_dir :

```
>>> dataset = load_dataset("pdffolder", data_dir="/path/to/folder")
```

Bạn cũng có thể sử dụng pdffolder để tải tập dữ liệu liên quan đến nhiều phần tách. Để làm như vậy, tập dữ thư mục nên có cấu trúc như sau:

```
thư mục/train/sơ yếu lý lịch/0001.pdf
thư mục/train/sơ yếu lý lịch/0002.pdf
thư mục/kiểm tra/hóa đơn/0001.pdf
thư mục/kiểm tra/hóa đơn/0002.pdf
```

> [!WARNING]
> Nếu tất cả các tệp PDF được chứa trong một thư mục duy nhất hoặc nếu chúng không ở cùng cấp độ cấu trúc thư mục, cột nhãn sẽ không được thêm tự động. Nếu bạn cần thì hãy đặt drop_labels=False explicitly.

Nếu có thông tin bổ sung mà bạn muốn đưa vào về tập dữ liệu của mình, như chú thích văn bản hoặc các hộp giới hạn, hãy thêm nó dưới dạng tệp siêu dữ liệu.csv trong thư mục của bạn. Điều này cho phép bạn bộ dữ liệu cho các tác vụ thị giác máy tính khác nhau như chú thích văn bản hoặc phát hiện đối tượng. bạn c cũng sử dụng tệp JSONL siêu dữ liệu.jsonl hoặc tệp Parquet siêu dữ liệu.parquet .

```
thư mục/train/metadata.csv
thư mục/tàu/0001.pdf
thư mục/tàu/0002.pdf
thư mục/tàu/0003.pdf
```

Tệp siêu dữ liệu.csv của bạn phải có trường file_name hoặc *_file_name liên kết các tệp PDF với siêu dữ liệu của họ:

```
tên_tệp,tính năng bổ sung
0001.pdf,Đây là giá trị đầu tiên của tính năng văn bản bạn đã thêm vào pdf của mình
0002.pdf,Đây là giá trị thứ hai của tính năng văn bản bạn đã thêm vào pdf của mình
0003.pdf,Đây là giá trị thứ ba của tính năng văn bản bạn đã thêm vào pdf của mình
```

hoặc sử dụng siêu dữ liệu.jsonl:

```
{"file_name": "0001.pdf", "additional_feature": "This is a first value of a text feature you added
{"file_name": "0002.pdf", "additional_feature": "This is a second value of a text feature you adde
{"file_name": "0003.pdf", "additional_feature": "This is a third value of a text feature you added
```

Ở đây file_name phải là tên của tệp PDF bên cạnh tệp siêu dữ liệu. Hơn
nói chung, nó phải là đường dẫn tương đối từ thư mục chứa siêu dữ liệu tới tệp PDF
tài liệu.

Có thể trỏ đến nhiều hơn một pdf trong mỗi hàng trong tập dữ liệu của bạn, ví dụ: nếu cả hai
đầu vào và đầu ra là pdf:

```
{"input_file_name": "0001.pdf", "output_file_name": "0001_output.pdf"}
{"input_file_name": "0002.pdf", "output_file_name": "0002_output.pdf"}
{"input_file_name": "0003.pdf", "output_file_name": "0003_output.pdf"}
```

Bạn cũng có thể xác định danh sách các tệp pdf. Trong trường hợp đó bạn cần đặt tên cho trường file_names
*_file_names . Đây là một ví dụ:

```
{"pdfs_file_names": ["0001_part1.pdf", "0001_part2.pdf"], "label": "urgent"}
{"pdfs_file_names": ["0002_part1.pdf", "0002_part2.pdf"], "label": "urgent"}
{"pdfs_file_names": ["0003_part1.pdf", "0002_part2.pdf"], "label": "normal"}
```

OCR (Optical character recognition)

Bộ dữ liệu OCR có văn bản chứa trong pdf. Một ví dụ về siêu dữ liệu.csv có thể trông giống như:

```
tên_tệp,văn bản
0001.pdf,Hóa đơn 1234 ngày 01/01/1970...
0002.pdf, Sơ yếu lý lịch kỹ sư phần mềm. Giáo dục: ...
0003.pdf,Chú ý là tất cả những gì bạn cần. Trừu tượng. ...
```

Tải tập dữ liệu bằng PdfFolder và nó sẽ tạo một cột văn bản cho chú thích pdf:

```
>>> dataset = load_dataset("pdffolder", data_dir="/path/to/folder", split="train")
>>> dataset[0]["text"]
“Hóa đơn 1234 ngày 01/01/1970...”
```

Tải tập dữ liệu lên Hub

Ví dụ: khi bạn đã tạo tập dữ liệu, bạn có thể chia sẻ tập dữ liệu đó với bằng cách sử dụng ôm mặt_hub. Đảm bảo bạn đã cài đặt thư viện ôm mặt_hub và bạn đã đăng nhập vào Hugging Face account (see the Upload with Python tutorial for more details).

Tải tập dữ liệu của bạn lên với ômface_hub.HfApi.upload_folder :

```
từ ôm mặt_hub nhập HfApi
api = HfApi()

api.upload_folder(
    folder_path="/path/to/local/dataset",
    repo_id="username/my-cool-dataset",
    repo_type="dataset",
)
```