Phương pháp tải

Phương pháp liệt kê và tải dữ liệu:

Bộ dữ liệudatasets.load_dataset

bộ dữ liệu.load_datasetdatasets.load_datasethttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/load.py#L1190[{"name": "path", "val": ": str"}, {"name": "name", "val": ": typing.Optional[str] = None"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[str, collections.abc.Sequence[str], collections.abc.Mapping[str, typing.Union[str, collections.abc.Sequence[str]]], NoneType] = None"}, {"name": "split", "val": ": typing.Union[str, datasets.splits.Split, list[str], list[datasets.splits.Split], NoneType] = None"}, {"name": "cache_dir", "val": ": typing.Optional[str] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "download_config", "val": ": typing.Optional[datasets.download.download_config.DownloadConfig] = None"}, {"name": "download_mode", "val": ": typing.Union[datasets.download.download_manager.DownloadMode, str, NoneType] = None"}, {"name": "verification_mode", "val": ": typing.Union[datasets.utils.info_utils.VerificationMode, str, NoneType] = None"}, {"name": "keep_in_memory", "val": ": typing.Optional[bool] = None"}, {"name": "save_infos", "val": ": bool = False"}, {"name": "revision", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name": "streaming", "val": ": bool = False"}, {"name": "num_proc", "val": ": typing.Optional[int] = None"}, {"name": "storage_options", "val": ": typing.Optional[dict] = None"}, {"name": "**config_kwargs", "val": ""}]- path (str) -- Đường dẫn hoặc tên của tập dữ liệu.

- if path is a dataset repository on the HF hub (list all available datasets with huggingface_hub.list_datasets)
  -> load the dataset from supported files in the repository (csv, json, parquet, etc.)
  ví dụ. 'username/dataset_name' , kho lưu trữ dữ liệu trên trung tâm HF chứa dữ liệu tập tin.
- nếu đường dẫn là một thư mục cục bộ
  -> load the dataset from supported files in the directory (csv, json, parquet, etc.)

ví dụ. './path/to/directory/with/my/csv/data' .

- nếu đường dẫn là tên của trình tạo tập dữ liệu và data_files hoặc data_dir được chỉ định (available builders are "json", "csv", "parquet", "arrow", "text", "xml", "webdataset", "imagefolder", "audiofolder", "videofolder")
  -> load the dataset from the files in data_files or data_dir
  ví dụ. 'sàn gỗ'.
- name (str , optional) --
  Xác định tên của cấu hình tập dữ liệu.
- data_dir (str , optional) --
  Xác định data_dir của cấu hình tập dữ liệu. Nếu được chỉ định cho các trình tạo chung (csv, text etc.) or the Hub datasets and data_files is None ,
  the behavior is equal to passing os.path.join(data_dir, **) as data_files to reference tất cả các tập tin trong một thư mục.
- data_files (str or Sequence or Mapping , optional) --
  Path(s) to source data file(s).
- split (Split or str) --
  Phân chia dữ liệu nào để tải.
  If None , will return a dict with all splits (typically datasets.Split.TRAIN and datasets.Split.TEST).
  Nếu được, sẽ trả về một Tập dữ liệu duy nhất.
  Việc phân chia có thể được kết hợp và chỉ định giống như trong tập dữ liệu tensorflow.
- cache_dir (str , optional) --
  Thư mục để đọc/ghi dữ liệu. Mặc định là "~/.cache/huggingface/datasets" .
- features (Features , optional) --
  Đặt loại tính năng sẽ sử dụng cho tập dữ liệu này.
- download_config (DownloadConfig, optional) --
  Thông số cấu hình tải xuống cụ thể.
- download_mode (DownloadMode or str , defaults to REUSE_DATASET_IF_EXISTS) --
  Chế độ tải xuống/tạo.
- verification_mode (VerificationMode or str , defaults to BASIC_CHECKS) --
  Chế độ xác minh xác định các bước kiểm tra để chạy trên tập dữ liệu đã tải xuống/đã xử lý information (checksums/size/splits/...).
- keep_in_memory (bool , defaults to None) --
  Có sao chép tập dữ liệu vào bộ nhớ hay không. Nếu Không , tập dữ liệu
  sẽ không được sao chép trong bộ nhớ trừ khi được bật rõ ràng bằng cách cài đặt

bộ dữ liệu.config.IN_MEMORY_MAX_SIZE thành
khác không. Xem thêm chi tiết trong phần cải thiện hiệu suất.

- revision (Version or str , optional) --
  Phiên bản của tập dữ liệu cần tải.
  Vì các bộ dữ liệu có kho lưu trữ git riêng trên Trung tâm bộ dữ liệu nên phiên bản mặc định là "chính"
  tương ứng với nhánh "chính" của họ.
  Bạn có thể chỉ định một phiên bản khác với phiên bản "chính" mặc định bằng cách sử dụng SHA hoặc gi
  thẻ của kho lưu trữ dữ liệu.
- token (str or bool , optional) --
  Chuỗi hoặc boolean tùy chọn để sử dụng làm mã thông báo Bearer cho các tệp từ xa trên Trung tâm bộ
  Nếu True hoặc không được chỉ định, sẽ nhận được mã thông báo từ "~/.huggingface" .
- streaming (bool , defaults to False) --
  Nếu được đặt thành True , không tải xuống tệp dữ liệu. Thay vào đó, nó truyền dữ liệu dần dần
  trong khi
  lặp lại trên tập dữ liệu. Thay vào đó, IterableDataset hoặc IterableDatasetDict được trả về trong
  trường hợp này.
  Lưu ý rằng tính năng phát trực tuyến hoạt động đối với các tập dữ liệu sử dụng định dạng dữ liệu hỗ trợ
  như txt, csv, jsonl chẳng hạn.
  Các tập tin Json có thể được tải xuống hoàn toàn. Cũng phát trực tuyến từ các tệp zip hoặc gzip từ xa là
  được hỗ trợ nhưng các định dạng nén khác
  như rar và xz chưa được hỗ trợ. Định dạng tgz không cho phép phát trực tuyến.
- num_proc (int , optional, defaults to None) --
  Số lượng quy trình khi tải xuống và tạo tập dữ liệu cục bộ.
  Đa xử lý bị tắt theo mặc định.
- storage_options (dict , optional, defaults to None) --
  Thực nghiệm. Các cặp khóa/giá trị sẽ được chuyển đến phần phụ trợ hệ thống tệp tập dữ liệu, nếu có.
- **config_kwargs (additional keyword arguments) --
  Đối số từ khóa được chuyển đến BuilderConfig
  và được sử dụng trong DatasetBuilder.0Dataset hoặc DatasetDict- nếu phần chia không phải là None : t
  yêu cầu,
- nếu phần tách là Không có, một DatasetDict với mỗi phần tách.

or IterableDataset or IterableDatasetDict: if streaming=True

- nếu phần chia không phải là None , tập dữ liệu sẽ được yêu cầu
- nếu phần tách là Không có, sẽ có ~datasets.streaming.IterableDatasetDict với mỗi phần chia.

Tải tập dữ liệu từ Hugging Face Hub hoặc tập dữ liệu cục bộ.

Bạn có thể tìm thấy danh sách các bộ dữ liệu trên Hub hoặc với ôm mặt_hub.list_datasets.

A dataset is a directory that contains some data files in generic formats (JSON, CSV, Parquet, etc.) and possibly
in a generic structure (Webdataset, ImageFolder, AudioFolder, VideoFolder, etc.)

Chức năng này thực hiện những điều sau đây:

1. Tải trình tạo tập dữ liệu:
   - Tìm định dạng dữ liệu phổ biến nhất trong tập dữ liệu và chọn trình tạo liên quan của nó (JSON, CSV, Parquet, Webdataset, ImageFolder, AudioFolder, etc.)
   - Find which file goes into which split (e.g. train/test) based on file and directory names hoặc trên cấu hình YAML
   - Cũng có thể chỉ định data_files theo cách thủ công và sử dụng trình tạo tập dữ liệu nào (e.g. "parquet").
2. Chạy trình tạo tập dữ liệu:

   Trong trường hợp tổng quát:
   - Tải xuống các tệp dữ liệu từ tập dữ liệu nếu chúng chưa có sẵn tại địa phương hoặc được lưu vào bộ nhớ đệm.
   - Xử lý và lưu trữ tập dữ liệu trong các bảng Mũi tên đã gõ để lưu vào bộ nhớ đệm.

     Bảng mũi tên là các bảng được định kiểu, dài tùy ý, có thể lưu trữ các đối tượng lồng nhau và được ánh xạ tới các loại chung chung numpy/pandas/python.

     Chúng có thể được truy cập trực tiếp từ đĩa, được tải vào RAM hoặc thậm chí được truyền trực tuyến web.

   Trong trường hợp phát trực tuyến:
   - Không tải xuống hoặc lưu vào bộ nhớ đệm bất cứ thứ gì. Thay vào đó, tập dữ liệu được tải một cách được phát trực tiếp khi lặp lại trên đó.
3. Return a dataset built from the requested splits in split (default: all).

Ví dụ:

Tải tập dữ liệu từ Hugging Face Hub:

```
>>> from datasets import load_dataset
>>> ds = load_dataset('cornell-movie-review-data/rotten_tomatoes', split='train')

# Load a subset or dataset configuration (here 'sst2')
>>> from datasets import load_dataset
>>> ds = load_dataset('nyu-mll/glue', 'sst2', split='train')

# Ánh xạ thủ công các tệp dữ liệu thành các phần tách
>>> data_files = {'train': 'train.csv', 'test': 'test.csv'}
>>> ds = load_dataset('namespace/your_dataset_name', data_files=data_files)

# Chọn thủ công thư mục để tải
>>> ds = load_dataset('namespace/your_dataset_name', data_dir='folder_name')
```

Tải tập dữ liệu cục bộ:

```
# Tải tệp CSV
>>> from datasets import load_dataset
>>> ds = load_dataset('csv', data_files='path/to/local/my_dataset.csv')

# Tải tệp JSON
>>> from datasets import load_dataset
>>> ds = load_dataset('json', data_files='path/to/local/my_dataset.json')
```

Tải một IterableDataset:

```
>>> from datasets import load_dataset
>>> ds = load_dataset('cornell-movie-review-data/rotten_tomatoes', split='train', streaming=True)
```

Tải tập dữ liệu hình ảnh bằng trình tạo tập dữ liệu ImageFolder:

```
>>> from datasets import load_dataset
>>> ds = load_dataset('imagefolder', data_dir='/path/to/images', split='train')
```

bộ dữ liệu.load_from_diskdatasets.load_from_diskhttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/load.py#L1434[{"name": "dataset_path", "val": ": typing.Union[str, bytes, os.PathLike]"}, {"name": "keep_in_memory", "val": ": typing.Optional[bool] = None"}, {"name":

"storage_options", "val": ": typing.Optional[dict] = None"}]- dataset_path (path-like) --
Path (e.g. "dataset/train") or remote URI (e.g. "s3://my-bucket/dataset/train")
của thư mục Dataset hoặc DatasetDict nơi tập dữ liệu/dataset-dict sẽ được lưu trữ
được tải từ.

- keep_in_memory (bool , defaults to None) --
  Có sao chép tập dữ liệu vào bộ nhớ hay không. Nếu Không , tập dữ liệu
  sẽ không được sao chép trong bộ nhớ trừ khi được bật rõ ràng bằng cách cài đặt
  bộ dữ liệu.config.IN_MEMORY_MAX_SIZE thành
  khác không. Xem thêm chi tiết trong phần cải thiện hiệu suất.
- storage_options (dict , optional) --
  Các cặp khóa/giá trị sẽ được chuyển đến phần phụ trợ hệ thống tệp, nếu có.
  0Dataset hoặc DatasetDict- Nếu tập dữ liệu_path là đường dẫn của thư mục tập dữ liệu: tập dữ liệu
  được yêu cầu.
- Nếu tập dữ liệu_path là đường dẫn của thư mục tập dữ liệu dict, DatasetDict với mỗi phần tách.

Loads a dataset that was previously saved using save_to_disk() from a dataset directory, or
từ hệ thống tệp bằng cách sử dụng bất kỳ triển khai fsspec.spec.AbstractFileSystem nào.

Ví dụ:

```
>>> from datasets import load_from_disk
>>> ds = load_from_disk('path/to/dataset/directory')
```

bộ dữ liệu.load_dataset_builderdatasets.load_dataset_builderhttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/load.py#L1041[{"name": "path", "val": ": str"}, {"name": "name", "val": ": typing.Optional[str] = None"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[str, collections.abc.Sequence[str], collections.abc.Mapping[str, typing.Union[str, collections.abc.Sequence[str]]], NoneType] = None"}, {"name": "cache_dir", "val": ": typing.Optional[str] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "download_config", "val": ": typing.Optional[datasets.download.download_config.DownloadConfig] = None"}, {"name": "download_mode", "val": ": typing.Union[datasets.download.download_manager.DownloadMode, str, NoneType] = None"}, {"name": "revision", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name":

"storage_options", "val": ": typing.Optional[dict] = None"}, {"name": "**config_kwargs", "val": ""}]- path (str) --
Đường dẫn hoặc tên của tập dữ liệu.

- if path is a dataset repository on the HF hub (list all available datasets with huggingface_hub.list_datasets)
  -> load the dataset builder from supported files in the repository (csv, json, parquet, etc.)
  ví dụ. 'username/dataset_name' , kho lưu trữ dữ liệu trên trung tâm HF chứa dữ liệu tập tin.
- nếu đường dẫn là một thư mục cục bộ
  -> load the dataset builder from supported files in the directory (csv, json, parquet, etc.)
  ví dụ. './path/to/directory/with/my/csv/data' .
- nếu đường dẫn là tên của trình tạo tập dữ liệu và data_files hoặc data_dir được chỉ định (available builders are "json", "csv", "parquet", "arrow", "text", "xml", "webdataset", "imagefolder", "audiofolder", "videofolder")
  -> load the dataset builder from the files in data_files or data_dir
  ví dụ. 'sàn gỗ'.
- name (str , optional) --
  Xác định tên của cấu hình tập dữ liệu.
- data_dir (str , optional) --
  Xác định data_dir của cấu hình tập dữ liệu. Nếu được chỉ định cho các trình tạo chung (csv, text etc.) or the Hub datasets and data_files is None ,
  the behavior is equal to passing os.path.join(data_dir, **) as data_files to reference tất cả các tập tin trong một thư mục.
- data_files (str or Sequence or Mapping , optional) --
  Path(s) to source data file(s).
- cache_dir (str , optional) --
  Thư mục để đọc/ghi dữ liệu. Mặc định là "~/.cache/huggingface/datasets" .
- features (Features, optional) --
  Đặt loại tính năng sẽ sử dụng cho tập dữ liệu này.
- download_config (DownloadConfig, optional) --
  Thông số cấu hình tải xuống cụ thể.
- download_mode (DownloadMode or str , defaults to REUSE_DATASET_IF_EXISTS) --
  Chế độ tải xuống/tạo.
- revision (Version or str , optional) --

Phiên bản của tập dữ liệu cần tải.
Vì các bộ dữ liệu có kho lưu trữ git riêng trên Trung tâm bộ dữ liệu nên phiên bản mặc định là "chính" tương ứng với nhánh "chính" của họ.
Bạn có thể chỉ định một phiên bản khác với phiên bản "chính" mặc định bằng cách sử dụng SHA hoặc gi thẻ của kho lưu trữ dữ liệu.

- token (str or bool , optional) --
  Chuỗi hoặc boolean tùy chọn để sử dụng làm mã thông báo Bearer cho các tệp từ xa trên Trung tâm bộ
  Nếu True hoặc không được chỉ định, sẽ nhận được mã thông báo từ "~/.huggingface" .
- storage_options (dict , optional, defaults to None) --
  Thực nghiệm. Các cặp khóa/giá trị sẽ được chuyển đến phần phụ trợ hệ thống tệp tập dữ liệu, nếu có.
- **config_kwargs (additional keyword arguments) --
  Đối số từ khóa được chuyển đến BuilderConfig
  và được sử dụng trong DatasetBuilder.0DatasetBuilder
  Tải trình tạo tập dữ liệu có thể được sử dụng để:
- Inspect general information that is required to build a dataset (cache directory, config, dataset info, features, data files, etc.)
- Tải xuống và chuẩn bị tập dữ liệu dưới dạng tệp Mũi tên trong bộ đệm
- Nhận tập dữ liệu phát trực tuyến mà không cần tải xuống hoặc lưu vào bộ nhớ đệm bất kỳ thứ gì

Bạn có thể tìm thấy danh sách các bộ dữ liệu trên Hub hoặc với ôm mặt_hub.list_datasets.

A dataset is a directory that contains some data files in generic formats (JSON, CSV, Parquet, etc.) and possibly
in a generic structure (Webdataset, ImageFolder, AudioFolder, VideoFolder, etc.)

Ví dụ:

```
>>> from datasets import load_dataset_builder
>>> ds_builder = load_dataset_builder('cornell-movie-review-data/rotten_tomatoes')
>>> ds_builder.info.features
{'label': ClassLabel(names=['neg', 'pos']),
 'text': Value('string')}
```

bộ dữ liệu.get_dataset_config_namesdatasets.get_dataset_config_nameshttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/inspect.py#L109[{"name": "path", "val": ": str"}, {"name": "revision", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = None"}, {"name": "download_config", "val": ":

typing.Optional[datasets.download.download_config.DownloadConfig] = None"}, {"name": "download_mode", "val": ": typing.Union[datasets.download.download_manager.DownloadMode, str, NoneType] = None"}, {"name": "data_files", "val": ": typing.Union[str, list, dict, NoneType] = None"}, {"name": "**download_kwargs", "val": ""}]- path (str) -- path to the dataset repository. Can be either:

- đường dẫn cục bộ tới thư mục tập dữ liệu chứa các tệp dữ liệu,
  ví dụ. './dataset/đội'
- a dataset identifier on the Hugging Face Hub (list all available datasets and ids with huggingface_hub.list_datasets),
  ví dụ. 'rajpurkar/squad' , 'nyu-mll/glue' hoặc``'openai/webtext'`
- revision (Union[str, datasets.Version] , optional) --
  Nếu được chỉ định, mô-đun tập dữ liệu sẽ được tải từ kho lưu trữ tập dữ liệu ở phiên bản này.
  Theo mặc định:
    nó được đặt thành phiên bản cục bộ của lib.
    nó cũng sẽ cố tải nó từ nhánh chính nếu nó không có sẵn ở phiên bản địa phương của lib.
      Việc chỉ định một phiên bản khác với phiên bản lib cục bộ của bạn có thể gây ra vấn đề tương thích.
- download_config (DownloadConfig, optional) --
  Thông số cấu hình tải xuống cụ thể.
- download_mode (DownloadMode or str , defaults to REUSE_DATASET_IF_EXISTS) --
  Chế độ tải xuống/tạo.
- data_files (Union[Dict, List, str] , optional) --
  Xác định data_files của cấu hình tập dữ liệu.
- **download_kwargs (additional keyword arguments) --
  Các thuộc tính tùy chọn cho DownloadConfig sẽ ghi đè các thuộc tính trong download_config nếu được cung cấp,
  ví dụ mã thông báo .0
  Lấy danh sách tên cấu hình có sẵn cho một tập dữ liệu cụ thể.

Ví dụ:

```
>>> from datasets import get_dataset_config_names
>>> get_dataset_config_names("nyu-mll/glue")
['cola',
 'sst2',
 'mrpc',
 'qqp',
 'stsb',
 'may mắn',
 'mnli_không khớp',
 'mnli_matched',
 'qnli',
 'rte',
 'wli',
 'ax']
```

bộ dữ liệu.get_dataset_infosdatasets.get_dataset_infoshttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/inspect.py#L42[{"name": "path", "val": ": str"}, {"name": "data_files", "val": ": typing.Union[str, list, dict, NoneType] = None"}, {"name": "download_config", "val": ": typing.Optional[datasets.download.download_config.DownloadConfig] = None"}, {"name": "download_mode", "val": ": typing.Union[datasets.download.download_manager.DownloadMode, str, NoneType] = None"}, {"name": "revision", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name": "**config_kwargs", "val": ""}]- path (str) -- path to the dataset repository. Can be either:

- đường dẫn cục bộ tới thư mục tập dữ liệu chứa các tệp dữ liệu,
  ví dụ. './dataset/đội'
- a dataset identifier on the Hugging Face Hub (list all available datasets and ids with huggingface_hub.list_datasets),
  ví dụ. 'rajpurkar/squad' , 'nyu-mll/glue' hoặc``'openai/webtext'`
- revision (Union[str, datasets.Version] , optional) --
  Nếu được chỉ định, mô-đun tập dữ liệu sẽ được tải từ kho lưu trữ tập dữ liệu ở phiên bản này.
  Theo mặc định:
  nó được đặt thành phiên bản cục bộ của lib.
  nó cũng sẽ cố tải nó từ nhánh chính nếu nó không có sẵn ở phiên bản địa phương của lib.
  Việc chỉ định một phiên bản khác với phiên bản lib cục bộ của bạn có thể gây ra

vấn đề tương thích.

- download_config (DownloadConfig, optional) --
  Thông số cấu hình tải xuống cụ thể.
- download_mode (DownloadMode or str , defaults to REUSE_DATASET_IF_EXISTS) --
  Chế độ tải xuống/tạo.
- data_files (Union[Dict, List, str] , optional) --
  Xác định data_files của cấu hình tập dữ liệu.
- token (str or bool , optional) --
  Chuỗi hoặc boolean tùy chọn để sử dụng làm mã thông báo Bearer cho các tệp từ xa trên Trung tâm bộ
  Nếu True hoặc không được chỉ định, sẽ nhận được mã thông báo từ "~/.huggingface" .
- **config_kwargs (additional keyword arguments) --
  Các thuộc tính tùy chọn cho lớp trình tạo sẽ ghi đè các thuộc tính nếu được cung cấp.0
  Nhận thông tin meta về tập dữ liệu, được trả về dưới dạng tên cấu hình ánh xạ chính tả cho
  Tập dữ liệuInfoDict.

Ví dụ:

```
>>> from datasets import get_dataset_infos
>>> get_dataset_infos('cornell-movie-review-data/rotten_tomatoes')
{'default': DatasetInfo(description="Movie Review Dataset.
  là tập dữ liệu chứa 5.331 kết quả dương tính và 5.331 kết quả âm tính được xử lý
ences from Rotten Tomatoes movie reviews...), ...}
```

bộ dữ liệu.get_dataset_split_namesdatasets.get_dataset_split_nameshttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/inspect.py#L298[{"name": "path", "val": ": str"}, {"name": "config_name", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[str, collections.abc.Sequence[str], collections.abc.Mapping[str, typing.Union[str, collections.abc.Sequence[str]]], NoneType] = None"}, {"name": "download_config", "val": ": typing.Optional[datasets.download.download_config.DownloadConfig] = None"}, {"name": "download_mode", "val": ": typing.Union[datasets.download.download_manager.DownloadMode, str, NoneType] = None"}, {"name": "revision", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name": "**config_kwargs", "val": ""}]- path (str) -- path to the dataset repository. Can be either:

- đường dẫn cục bộ tới thư mục tập dữ liệu chứa các tệp dữ liệu,

ví dụ. './dataset/đội'

- a dataset identifier on the Hugging Face Hub (list all available datasets and ids with huggingface_hub.list_datasets),
  ví dụ. 'rajpurkar/squad' , 'nyu-mll/glue' hoặc``'openai/webtext'`
- config_name (str , optional) --
  Xác định tên của cấu hình tập dữ liệu.
- data_files (str or Sequence or Mapping , optional) --
  Path(s) to source data file(s).
- download_config (DownloadConfig, optional) --
  Thông số cấu hình tải xuống cụ thể.
- download_mode (DownloadMode or str , defaults to REUSE_DATASET_IF_EXISTS) --
  Chế độ tải xuống/tạo.
- revision (Version or str , optional) --
  Phiên bản của tập dữ liệu cần tải.
  Vì các bộ dữ liệu có kho lưu trữ git riêng trên Trung tâm bộ dữ liệu nên phiên bản mặc định là "chính"
  tương ứng với nhánh "chính" của họ.
  Bạn có thể chỉ định một phiên bản khác với phiên bản "chính" mặc định bằng cách sử dụng SHA hoặc gi
  thẻ của kho lưu trữ dữ liệu.
- token (str or bool , optional) --
  Chuỗi hoặc boolean tùy chọn để sử dụng làm mã thông báo Bearer cho các tệp từ xa trên Trung tâm bộ
  Nếu True hoặc không được chỉ định, sẽ nhận được mã thông báo từ "~/.huggingface" .
- **config_kwargs (additional keyword arguments) --
  Các thuộc tính tùy chọn cho lớp trình tạo sẽ ghi đè các thuộc tính nếu được cung cấp.0
  Nhận danh sách các phần tách có sẵn cho một cấu hình và tập dữ liệu cụ thể.

Ví dụ:

```
>>> from datasets import get_dataset_split_names
>>> get_dataset_split_names('cornell-movie-review-data/rotten_tomatoes')
['train', 'validation', 'test']
```

Từ tập tin

Các cấu hình dùng để tải file dữ liệu.
Chúng được sử dụng khi tải các tệp cục bộ hoặc kho lưu trữ dữ liệu:

- local files: load_dataset("parquet", data_dir="path/to/data/dir")
- dataset repository: load_dataset("allenai/c4")

Bạn có thể chuyển đối số cho Load_dataset để định cấu hình tải dữ liệu.
Ví dụ: bạn có thể chỉ định tham số sep để xác định CsvConfig được sử dụng để tải dữ liệu:

```
load_dataset("csv", data_dir="path/to/data/dir", sep="\t")
```

Textdatasets.packaged_modules.text.TextConfig

lớp học
bộ dữ liệu.packaged_modules.text.TextConfigdatasets.packaged_modules.text.TextConfighttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/text/text.py#L17[{"name": "name", "val": ": str = 'default'"}, {"name": "version", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = 0.0.0"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[datasets.data_files.DataFilesDict, datasets.data_files.DataFilesPatternsDict, NoneType] = None"}, {"name": "description", "val": ": typing.Optional[str] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "encoding", "val": ": str = 'utf-8'"}, {"name": "encoding_errors", "val": ": typing.Optional[str] = None"}, {"name": "chunksize", "val": ": int = 10485760"}, {"name": "keep_linebreaks", "val": ": bool = False"}, {"name": "sample_by", "val": ": str = 'line'"}]
BuilderConfig cho các tệp văn bản.

lớp học
bộ dữ liệu.packaged_modules.text.Textdatasets.packaged_modules.text.Texthttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/text/text.py#L28[{"name": "cache_dir", "val": ": typing.Optional[str] = None"}, {"name": "dataset_name", "val": ": typing.Optional[str] = None"}, {"name": "config_name", "val": ": typing.Optional[str] = None"}, {"name": "hash", "val": ": typing.Optional[str] = None"}, {"name": "base_path", "val": ": typing.Optional[str] = None"}, {"name": "info", "val": ": typing.Optional[datasets.info.DatasetInfo] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name": "repo_id", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[str, list, dict, datasets.data_files.DataFilesDict, NoneType] = None"}, {"name": "data_dir", "val": ":

typing.Optional[str] = None"}, {"name": "storage_options", "val": ": typing.Optional[dict] = None"}, {"name": "writer_batch_size", "val": ": typing.Optional[int] = None"}, {"name": "**config_kwargs", "val": ""}]

CSVdatasets.packaged_modules.csv.CsvConfig

lớp học
bộ dữ liệu.packaged_modules.csv.CsvConfigdatasets.packaged_modules.csv.CsvConfighttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/csv/csv.py#L25[{"name": "name", "val": ": str = 'default'"}, {"name": "version", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = 0.0.0"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[datasets.data_files.DataFilesDict, datasets.data_files.DataFilesPatternsDict, NoneType] = None"}, {"name": "description", "val": ": typing.Optional[str] = None"}, {"name": "sep", "val": ": str = ','"}, {"name": "delimiter", "val": ": typing.Optional[str] = None"}, {"name": "header", "val": ": typing.Union[int, list[int], str, NoneType] = 'infer'"}, {"name": "names", "val": ": typing.Optional[list[str]] = None"}, {"name": "column_names", "val": ": typing.Optional[list[str]] = None"}, {"name": "index_col", "val": ": typing.Union[int, str, list[int], list[str], NoneType] = None"}, {"name": "usecols", "val": ": typing.Union[list[int], list[str], NoneType] = None"}, {"name": "prefix", "val": ": typing.Optional[str] = None"}, {"name": "mangle_dupe_cols", "val": ": bool = True"}, {"name": "engine", "val": ": typing.Optional[typing.Literal['c', 'python', 'pyarrow']] = None"}, {"name": "converters", "val": ": dict = None"}, {"name": "true_values", "val": ": typing.Optional[list] = None"}, {"name": "false_values", "val": ": typing.Optional[list] = None"}, {"name": "skipinitialspace", "val": ": bool = False"}, {"name": "skiprows", "val": ": typing.Union[int, list[int], NoneType] = None"}, {"name": "nrows", "val": ": typing.Optional[int] = None"}, {"name": "na_values", "val": ": typing.Union[str, list[str], NoneType] = None"}, {"name": "keep_default_na", "val": ": bool = True"}, {"name": "na_filter", "val": ": bool = True"}, {"name": "verbose", "val": ": bool = False"}, {"name": "skip_blank_lines", "val": ": bool = True"}, {"name": "thousands", "val": ": typing.Optional[str] = None"}, {"name": "decimal", "val": ": str = '.'"}, {"name": "lineterminator", "val": ": typing.Optional[str] = None"}, {"name": "quotechar", "val": ": str = '"'"}, {"name": "quoting", "val": ": int = 0"}, {"name": "escapechar", "val": ": typing.Optional[str] = None"}, {"name": "comment", "val": ": typing.Optional[str] = None"}, {"name": "encoding", "val": ": typing.Optional[str] = None"}, {"name": "dialect", "val": ": typing.Optional[str] = None"}, {"name": "error_bad_lines", "val": ": bool = True"}, {"name": "warn_bad_lines", "val": ": bool = True"}, {"name": "skipfooter", "val": ": int = 0"}, {"name":

"doublequote", "val": ": bool = True"}, {"name": "memory_map", "val": ": bool = False"}, {"name": "float_precision", "val": ": typing.Optional[str] = None"}, {"name": "chunksize", "val": ": int = 10000"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "encoding_errors", "val": ": typing.Optional[str] = 'strict'"}, {"name": "on_bad_lines", "val": ": typing.Literal['error', 'warn', 'skip'] = 'error'"}, {"name": "date_format", "val": ": typing.Optional[str] = None"}]
BuilderConfig cho CSV.

lớp học
bộ dữ liệu.packaged_modules.csv.Csvdatasets.packaged_modules.csv.Csvhttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/csv/csv.py#L145[{"name": "cache_dir", "val": ": typing.Optional[str] = None"}, {"name": "dataset_name", "val": ": typing.Optional[str] = None"}, {"name": "config_name", "val": ": typing.Optional[str] = None"}, {"name": "hash", "val": ": typing.Optional[str] = None"}, {"name": "base_path", "val": ": typing.Optional[str] = None"}, {"name": "info", "val": ": typing.Optional[datasets.info.DatasetInfo] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name": "repo_id", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[str, list, dict, datasets.data_files.DataFilesDict, NoneType] = None"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "storage_options", "val": ": typing.Optional[dict] = None"}, {"name": "writer_batch_size", "val": ": typing.Optional[int] = None"}, {"name": "**config_kwargs", "val": ""}]

JSONdatasets.packaged_modules.json.JsonConfig

lớp học
bộ dữ liệu.packaged_modules.json.JsonConfigdatasets.packaged_modules.json.JsonConfighttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/json/json.py#L42[{"name": "name", "val": ": str = 'default'"}, {"name": "version", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = 0.0.0"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[datasets.data_files.DataFilesDict, datasets.data_files.DataFilesPatternsDict, NoneType] = None"}, {"name": "description", "val": ": typing.Optional[str] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "encoding", "val": ": str = 'utf-8'"}, {"name": "encoding_errors", "val": ": typing.Optional[str] = None"}, {"name": "field", "val": ": typing.Optional[str] = None"}, {"name": "use_threads", "val": ":

bool = True"}, {"name": "block_size", "val": ": typing.Optional[int] = None"}, {"name": "chunksize", "val": ": int = 10485760"}, {"name": "newlines_in_values", "val": ": typing.Optional[bool] = None"}]
BuilderConfig cho JSON.

lớp học
bộ dữ liệu.packaged_modules.json.Jsondatasets.packaged_modules.json.Jsonhttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/json/json.py#L58[{"name": "cache_dir", "val": ": typing.Optional[str] = None"}, {"name": "dataset_name", "val": ": typing.Optional[str] = None"}, {"name": "config_name", "val": ": typing.Optional[str] = None"}, {"name": "hash", "val": ": typing.Optional[str] = None"}, {"name": "base_path", "val": ": typing.Optional[str] = None"}, {"name": "info", "val": ": typing.Optional[datasets.info.DatasetInfo] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name": "repo_id", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[str, list, dict, datasets.data_files.DataFilesDict, NoneType] = None"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "storage_options", "val": ": typing.Optional[dict] = None"}, {"name": "writer_batch_size", "val": ": typing.Optional[int] = None"}, {"name": "**config_kwargs", "val": ""}]

XMLdatasets.packaged_modules.xml.XmlConfig

lớp học
bộ dữ liệu.packaged_modules.xml.XmlConfigdatasets.packaged_modules.xml.XmlConfighttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/xml/xml.py#L16[{"name": "name", "val": ": str = 'default'"}, {"name": "version", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = 0.0.0"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[datasets.data_files.DataFilesDict, datasets.data_files.DataFilesPatternsDict, NoneType] = None"}, {"name": "description", "val": ": typing.Optional[str] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "encoding", "val": ": str = 'utf-8'"}, {"name": "encoding_errors", "val": ": typing.Optional[str] = None"}]
BuilderConfig cho các tệp xml.

lớp học

bộ dữ liệu.packaged_modules.xml.Xmldatasets.packaged_modules.xml.Xmlhttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/xml/xml.py#L24[{"name": "cache_dir", "val": ": typing.Optional[str] = None"}, {"name": "dataset_name", "val": ": typing.Optional[str] = None"}, {"name": "config_name", "val": ": typing.Optional[str] = None"}, {"name": "hash", "val": ": typing.Optional[str] = None"}, {"name": "base_path", "val": ": typing.Optional[str] = None"}, {"name": "info", "val": ": typing.Optional[datasets.info.DatasetInfo] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name": "repo_id", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[str, list, dict, datasets.data_files.DataFilesDict, NoneType] = None"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "storage_options", "val": ": typing.Optional[dict] = None"}, {"name": "writer_batch_size", "val": ": typing.Optional[int] = None"}, {"name": "**config_kwargs", "val": ""}]

Parquetdatasets.packaged_modules.parquet.ParquetConfig

lớp học
bộ dữ liệu.packaged_modules.parquet.ParquetConfigdatasets.packaged_modules.parquet.ParquetConfighttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/parquet/parquet.py#L17[{"name": "name", "val": ": str = 'default'"}, {"name": "version", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = 0.0.0"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[datasets.data_files.DataFilesDict, datasets.data_files.DataFilesPatternsDict, NoneType] = None"}, {"name": "description", "val": ": typing.Optional[str] = None"}, {"name": "batch_size", "val": ": typing.Optional[int] = None"}, {"name": "columns", "val": ": typing.Optional[list[str]] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "filters", "val": ": typing.Union[pyarrow._compute.Expression, list[tuple], list[list[tuple]], NoneType] = None"}, {"name": "fragment_scan_options", "val": ": typing.Optional[pyarrow._dataset_parquet.ParquetFragmentScanOptions] = None"}, {"name": "on_bad_files", "val": ": typing.Literal['error', 'warn', 'skip'] = 'error'"}]- batch_size (int , optional) --
Kích thước của RecordBatches để lặp lại.
The default is the row group size (defined by the first row group).

- columns (list[str] , optional) --
  Danh sách các cột cần tải, những cột khác bị bỏ qua.
  Tất cả các cột được tải theo mặc định.
- features -- (Features , optional):
  Truyền dữ liệu tới các tính năng.
- filters (Union[pyarrow.dataset.Expression, list[tuple], list[list[tuple]]] , optional) --
  Chỉ trả về các hàng phù hợp với bộ lọc.
  Nếu có thể vị ngữ sẽ bị đẩy xuống để khai thác thông tin phân vùng
  hoặc siêu dữ liệu nội bộ được tìm thấy trong nguồn dữ liệu, ví dụ: Thống kê sàn gỗ.
  Mặt khác, hãy lọc các RecordBatches đã tải trước khi mang lại chúng.
- fragment_scan_options (pyarrow.dataset.ParquetFragmentScanOptions , optional) --
  Tùy chọn quét cụ thể cho các mảnh Parquet.
  Điều này đặc biệt hữu ích để cấu hình bộ đệm và bộ nhớ đệm.
- on_bad_files (Literal["error", "warn", "skip"] , optional, defaults to "error") --
  Specify what to do upon encountering a bad file (a file that can't be read). Allowed values là :
    'lỗi', đưa ra Ngoại lệ khi gặp phải tệp xấu.
    'cảnh báo', đưa ra cảnh báo khi gặp phải một tệp xấu và bỏ qua tệp đó.
    'skip', bỏ qua các file xấu mà không đưa ra hoặc cảnh báo khi gặp phải.
  0

BuilderConfig cho sàn gỗ.

Ví dụ:

Tải một tập hợp con các cột:

```
>>> ds = load_dataset(parquet_dataset_id, columns=["col_0", "col_1"])
```

Truyền dữ liệu và lọc dữ liệu hiệu quả, có thể bỏ qua toàn bộ tệp hoặc nhóm hàng:

```
>>> filters = [("col_0", "==", 0)]
>>> ds = load_dataset(parquet_dataset_id, streaming=True, filters=filters)
```

Increase the minimum request size when streaming from 32MiB (default) to 128MiB and bật tìm nạp trước:

```
>>> import pyarrow
>>> import pyarrow.dataset
>>> fragment_scan_options = pyarrow.dataset.ParquetFragmentScanOptions(
...cache_options=pyarrow.CacheOptions(
...prefetch_limit=1,
...range_size_limit=128 << 20
...),
... )
>>> ds = load_dataset(parquet_dataset_id, streaming=True, fragment_scan_options=fragment_scan_opti
```

lớp học
bộ dữ liệu.packaged_modules.parquet.Parquetdatasets.packaged_modules.parquet.Parquethtt
ps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/parquet/
parquet.py#L90[{"name": "cache_dir", "val": ": typing.Optional[str] = None"}, {"name":
"dataset_name", "val": ": typing.Optional[str] = None"}, {"name": "config_name", "val": ":
typing.Optional[str] = None"}, {"name": "hash", "val": ": typing.Optional[str] = None"}, {"name":
"base_path", "val": ": typing.Optional[str] = None"}, {"name": "info", "val": ":
typing.Optional[datasets.info.DatasetInfo] = None"}, {"name": "features", "val": ":
typing.Optional[datasets.features.features.Features] = None"}, {"name": "token", "val": ":
typing.Union[bool, str, NoneType] = None"}, {"name": "repo_id", "val": ": typing.Optional[str] =
None"}, {"name": "data_files", "val": ": typing.Union[str, list, dict,
datasets.data_files.DataFilesDict, NoneType] = None"}, {"name": "data_dir", "val": ":
typing.Optional[str] = None"}, {"name": "storage_options", "val": ": typing.Optional[dict] =
None"}, {"name": "writer_batch_size", "val": ": typing.Optional[int] = None"}, {"name":
"**config_kwargs", "val": ""}]

Arrowdatasets.packaged_modules.arrow.ArrowConfig

lớp học
bộ dữ liệu.packaged_modules.arrow.ArrowConfigdatasets.packaged_modules.arrow.ArrowConfi
ghttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/arrow/
arrow.py#L15[{"name": "name", "val": ": str = 'default'"}, {"name": "version", "val": ":
typing.Union[datasets.utils.version.Version, str, NoneType] = 0.0.0"}, {"name": "data_dir", "val":
": typing.Optional[str] = None"}, {"name": "data_files", "val": ":
typing.Union[datasets.data_files.DataFilesDict, datasets.data_files.DataFilesPatternsDict,
NoneType] = None"}, {"name": "description", "val": ": typing.Optional[str] = None"}, {"name":

"features", "val": ": typing.Optional[datasets.features.features.Features] = None"}]
BuilderConfig cho Mũi tên.

lớp học
bộ dữ liệu.packaged_modules.arrow.Arrowdatasets.packaged_modules.arrow.Arrowhttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/arrow/arrow.py#L24[{"name": "cache_dir", "val": ": typing.Optional[str] = None"}, {"name": "dataset_name", "val": ": typing.Optional[str] = None"}, {"name": "config_name", "val": ": typing.Optional[str] = None"}, {"name": "hash", "val": ": typing.Optional[str] = None"}, {"name": "base_path", "val": ": typing.Optional[str] = None"}, {"name": "info", "val": ": typing.Optional[datasets.info.DatasetInfo] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name": "repo_id", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[str, list, dict, datasets.data_files.DataFilesDict, NoneType] = None"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "storage_options", "val": ": typing.Optional[dict] = None"}, {"name": "writer_batch_size", "val": ": typing.Optional[int] = None"}, {"name": "**config_kwargs", "val": ""}]

SQLdatasets.packaged_modules.sql.SqlConfig

lớp học
bộ dữ liệu.packaged_modules.sql.SqlConfigdatasets.packaged_modules.sql.SqlConfighttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/sql/sql.py#L24[{"name": "name", "val": ": str = 'default'"}, {"name": "version", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = 0.0.0"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[datasets.data_files.DataFilesDict, datasets.data_files.DataFilesPatternsDict, NoneType] = None"}, {"name": "description", "val": ": typing.Optional[str] = None"}, {"name": "sql", "val": ": typing.Union[str, ForwardRef('sqlalchemy.sql.Selectable')] = None"}, {"name": "con", "val": ": typing.Union[str, ForwardRef('sqlalchemy.engine.Connection'), ForwardRef('sqlalchemy.engine.Engine'), ForwardRef('sqlite3.Connection')] = None"}, {"name": "index_col", "val": ": typing.Union[str, list[str], NoneType] = None"}, {"name": "coerce_float", "val": ": bool = True"}, {"name": "params", "val": ": typing.Union[list, tuple, dict, NoneType] = None"}, {"name": "parse_dates", "val": ": typing.Union[list, dict, NoneType] = None"}, {"name": "columns", "val": ": typing.Optional[list[str]] = None"}, {"name": "chunksize", "val": ":

typing.Optional[int] = 10000"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}]
BuilderConfig cho SQL.

lớp học
bộ dữ liệu.packaged_modules.sql.Sqldatasets.packaged_modules.sql.Sqlhttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/sql/sql.py#L91[{"name": "cache_dir", "val": ": typing.Optional[str] = None"}, {"name": "dataset_name", "val": ": typing.Optional[str] = None"}, {"name": "config_name", "val": ": typing.Optional[str] = None"}, {"name": "hash", "val": ": typing.Optional[str] = None"}, {"name": "base_path", "val": ": typing.Optional[str] = None"}, {"name": "info", "val": ": typing.Optional[datasets.info.DatasetInfo] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name": "repo_id", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[str, list, dict, datasets.data_files.DataFilesDict, NoneType] = None"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "storage_options", "val": ": typing.Optional[dict] = None"}, {"name": "writer_batch_size", "val": ": typing.Optional[int] = None"}, {"name": "**config_kwargs", "val": ""}]

Imagesdatasets.packaged_modules.imagefolder.ImageFolderConfig

lớp học
bộ dữ liệu.packaged_modules.imagefolder.ImageFolderConfigdatasets.packaged_modules.imagefolder.ImageFolderConfighttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/imagefolder/imagefolder.py#L9[{"name": "name", "val": ": str = 'default'"}, {"name": "version", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = 0.0.0"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[datasets.data_files.DataFilesDict, datasets.data_files.DataFilesPatternsDict, NoneType] = None"}, {"name": "description", "val": ": typing.Optional[str] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "drop_labels", "val": ": bool = None"}, {"name": "drop_metadata", "val": ": bool = None"}, {"name": "metadata_filenames", "val": ": list = None"}, {"name": "filters", "val": ": typing.Union[pyarrow._compute.Expression, list[tuple], list[list[tuple]], NoneType] = None"}]
BuilderConfig cho ImageFolder.

lớp học
bộ dữ liệu.packaged_modules.imagefolder.ImageFolderdatasets.packaged_modules.imagefolder.ImageFolderhttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/imagefolder/imagefolder.py#L19[{"name": "cache_dir", "val": ": typing.Optional[str] = None"}, {"name": "dataset_name", "val": ": typing.Optional[str] = None"}, {"name": "config_name", "val": ": typing.Optional[str] = None"}, {"name": "hash", "val": ": typing.Optional[str] = None"}, {"name": "base_path", "val": ": typing.Optional[str] = None"}, {"name": "info", "val": ": typing.Optional[datasets.info.DatasetInfo] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name": "repo_id", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[str, list, dict, datasets.data_files.DataFilesDict, NoneType] = None"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "storage_options", "val": ": typing.Optional[dict] = None"}, {"name": "writer_batch_size", "val": ": typing.Optional[int] = None"}, {"name": "**config_kwargs", "val": ""}]

Audiodatasets.packaged_modules.audiofolder.AudioFolderConfig

lớp học
bộ dữ liệu.packaged_modules.audiofolder.AudioFolderConfigdatasets.packaged_modules.audiothư mục.AudioFolderConfighttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/audiofolder/audiofolder.py#L9[{"name": "name", "val": ": str = 'default'"}, {"name": "version", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = 0.0.0"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[datasets.data_files.DataFilesDict, datasets.data_files.DataFilesPatternsDict, NoneType] = None"}, {"name": "description", "val": ": typing.Optional[str] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "drop_labels", "val": ": bool = None"}, {"name": "drop_metadata", "val": ": bool = None"}, {"name": "metadata_filenames", "val": ": list = None"}, {"name": "filters", "val": ": typing.Union[pyarrow._compute.Expression, list[tuple], list[list[tuple]], NoneType] = None"}]
Cấu hình Builder cho AudioFolder.

lớp học
bộ dữ liệu.packaged_modules.audiofolder.AudioFolderdatasets.packaged_modules.audiofolder.AudioFolderhttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/

packaged_modules/audiofolder/audiofolder.py#L19[{"name": "cache_dir", "val": ": typing.Optional[str] = None"}, {"name": "dataset_name", "val": ": typing.Optional[str] = None"}, {"name": "config_name", "val": ": typing.Optional[str] = None"}, {"name": "hash", "val": ": typing.Optional[str] = None"}, {"name": "base_path", "val": ": typing.Optional[str] = None"}, {"name": "info", "val": ": typing.Optional[datasets.info.DatasetInfo] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name": "repo_id", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[str, list, dict, datasets.data_files.DataFilesDict, NoneType] = None"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "storage_options", "val": ": typing.Optional[dict] = None"}, {"name": "writer_batch_size", "val": ": typing.Optional[int] = None"}, {"name": "**config_kwargs", "val": ""}]

Videosdatasets.packaged_modules.videofolder.VideoFolderConfig

lớp học
bộ dữ liệu.packaged_modules.videofolder.VideoFolderConfigdatasets.packaged_modules.video thư mục.VideoFolderConfighttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/videofolder/videofolder.py#L9[{"name": "name", "val": ": str = 'default'"}, {"name": "version", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = 0.0.0"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[datasets.data_files.DataFilesDict, datasets.data_files.DataFilesPatternsDict, NoneType] = None"}, {"name": "description", "val": ": typing.Optional[str] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "drop_labels", "val": ": bool = None"}, {"name": "drop_metadata", "val": ": bool = None"}, {"name": "metadata_filenames", "val": ": list = None"}, {"name": "filters", "val": ": typing.Union[pyarrow._compute.Expression, list[tuple], list[list[tuple]], NoneType] = None"}]
BuilderConfig cho ImageFolder.

lớp học
bộ dữ liệu.packaged_modules.videofolder.VideoFolderdatasets.packaged_modules.videofolder.VideoFolderhttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/videofolder/videofolder.py#L19[{"name": "cache_dir", "val": ": typing.Optional[str] = None"}, {"name": "dataset_name", "val": ": typing.Optional[str] = None"}, {"name": "config_name", "val": ": typing.Optional[str] = None"}, {"name": "hash", "val": ":

typing.Optional[str] = None"}, {"name": "base_path", "val": ": typing.Optional[str] = None"}, {"name": "info", "val": ": typing.Optional[datasets.info.DatasetInfo] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name": "repo_id", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[str, list, dict, datasets.data_files.DataFilesDict, NoneType] = None"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "storage_options", "val": ": typing.Optional[dict] = None"}, {"name": "writer_batch_size", "val": ": typing.Optional[int] = None"}, {"name": "**config_kwargs", "val": ""}]

HDF5datasets.packaged_modules.hdf5.HDF5Config

lớp học
bộ dữ liệu.packaged_modules.hdf5.HDF5Configdatasets.packaged_modules.hdf5.HDF5Config https://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/hdf5/hdf5.py#L33[{"name": "name", "val": ": str = 'default'"}, {"name": "version", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = 0.0.0"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[datasets.data_files.DataFilesDict, datasets.data_files.DataFilesPatternsDict, NoneType] = None"}, {"name": "description", "val": ": typing.Optional[str] = None"}, {"name": "batch_size", "val": ": typing.Optional[int] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}]
BuilderConfig cho HDF5.

lớp học
bộ dữ liệu.packaged_modules.hdf5.HDF5datasets.packaged_modules.hdf5.HDF5https://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/hdf5/hdf5.py#L40[{"name": "cache_dir", "val": ": typing.Optional[str] = None"}, {"name": "dataset_name", "val": ": typing.Optional[str] = None"}, {"name": "config_name", "val": ": typing.Optional[str] = None"}, {"name": "hash", "val": ": typing.Optional[str] = None"}, {"name": "base_path", "val": ": typing.Optional[str] = None"}, {"name": "info", "val": ": typing.Optional[datasets.info.DatasetInfo] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name": "repo_id", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[str, list, dict, datasets.data_files.DataFilesDict, NoneType] = None"}, {"name": "data_dir", "val": ":

typing.Optional[str] = None"}, {"name": "storage_options", "val": ": typing.Optional[dict] = None"}, {"name": "writer_batch_size", "val": ": typing.Optional[int] = None"}, {"name": "**config_kwargs", "val": ""}]
ArrowBasedBuilder chuyển đổi các tệp HDF5 thành bảng Mũi tên bằng cách sử dụng các loại tiện ích mở rộn

Pdfdatasets.packaged_modules.pdffolder.PdfFolderConfig

lớp học
bộ dữ liệu.packaged_modules.pdffolder.PdfFolderConfigdatasets.packaged_modules.pdffolder.PdfFolderConfighttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/pdffolder/pdffolder.py#L9[{"name": "name", "val": ": str = 'default'"}, {"name": "version", "val": ": typing.Union[datasets.utils.version.Version, str, NoneType] = 0.0.0"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[datasets.data_files.DataFilesDict, datasets.data_files.DataFilesPatternsDict, NoneType] = None"}, {"name": "description", "val": ": typing.Optional[str] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "drop_labels", "val": ": bool = None"}, {"name": "drop_metadata", "val": ": bool = None"}, {"name": "metadata_filenames", "val": ": list = None"}, {"name": "filters", "val": ": typing.Union[pyarrow._compute.Expression, list[tuple], list[list[tuple]], NoneType] = None"}]
BuilderConfig cho ImageFolder.

lớp học
bộ dữ liệu.packaged_modules.pdffolder.PdfFolderdatasets.packaged_modules.pdffolder.PdfFolderhttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/pdffolder/pdffolder.py#L19[{"name": "cache_dir", "val": ": typing.Optional[str] = None"}, {"name": "dataset_name", "val": ": typing.Optional[str] = None"}, {"name": "config_name", "val": ": typing.Optional[str] = None"}, {"name": "hash", "val": ": typing.Optional[str] = None"}, {"name": "base_path", "val": ": typing.Optional[str] = None"}, {"name": "info", "val": ": typing.Optional[datasets.info.DatasetInfo] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name": "repo_id", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[str, list, dict, datasets.data_files.DataFilesDict, NoneType] = None"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "storage_options", "val": ": typing.Optional[dict] = None"}, {"name": "writer_batch_size", "val": ": typing.Optional[int] = None"}, {"name": "**config_kwargs", "val": ""}]

WebDatasetdatasets.packaged_modules.webdataset.WebTập dữ liệu

lớp học

bộ dữ liệu.packaged_modules.webdataset.WebDatasetdatasets.packaged_modules.webdataset.WebDatasethttps://github.com/huggingface/datasets/blob/4.2.0/src/datasets/packaged_modules/webdataset/webdataset.py#L19[{"name": "cache_dir", "val": ": typing.Optional[str] = None"}, {"name": "dataset_name", "val": ": typing.Optional[str] = None"}, {"name": "config_name", "val": ": typing.Optional[str] = None"}, {"name": "hash", "val": ": typing.Optional[str] = None"}, {"name": "base_path", "val": ": typing.Optional[str] = None"}, {"name": "info", "val": ": typing.Optional[datasets.info.DatasetInfo] = None"}, {"name": "features", "val": ": typing.Optional[datasets.features.features.Features] = None"}, {"name": "token", "val": ": typing.Union[bool, str, NoneType] = None"}, {"name": "repo_id", "val": ": typing.Optional[str] = None"}, {"name": "data_files", "val": ": typing.Union[str, list, dict, datasets.data_files.DataFilesDict, NoneType] = None"}, {"name": "data_dir", "val": ": typing.Optional[str] = None"}, {"name": "storage_options", "val": ": typing.Optional[dict] = None"}, {"name": "writer_batch_size", "val": ": typing.Optional[int] = None"}, {"name": "**config_kwargs", "val": ""}]