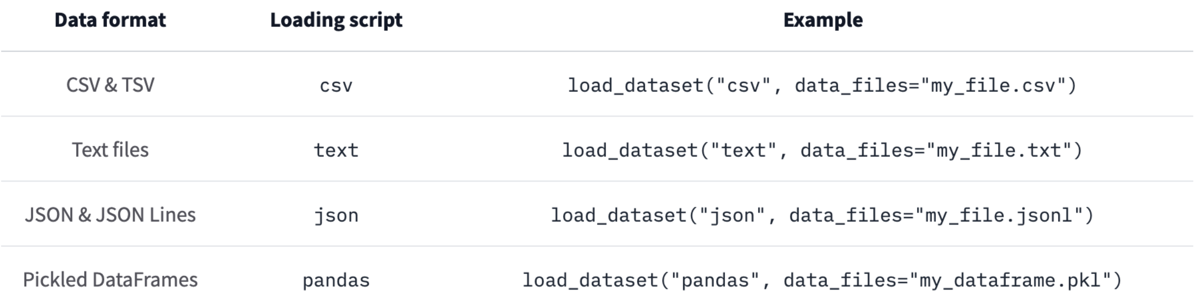
The Datasets Library

Important Agenda in this chapter:

- What do you do when your dataset is not on the Hub?
- How can you slice and dice a dataset? (And what if you *really* need to use Pandas?)
- What do you do when your dataset is huge and will melt your laptop’s RAM?
- What the heck are “memory mapping” and Apache Arrow?
- How can you create your own dataset and push it to the Hub?

What if my dataset isn't on the Hub?

🤗 Datasets provides loading scripts to handle the loading of local and remote datasets. It supports several common data formats, such as:



Data format	Loading script	Example
CSV & TSV	`csv`	`load_dataset("csv", data_files="my_file.csv")`
Text files	`text`	`load_dataset("text", data_files="my_file.txt")`
JSON & JSON Lines	`json`	`load_dataset("json", data_files="my_file.jsonl")`
Pickled DataFrames	`pandas`	`load_dataset("pandas", data_files="my_dataframe.pkl")`

Some important notes on the load_dataset method:

1. It can take input a local file in above mentioned formats.
2. It can consume a zip file directly

3. It can consume a remote url and load the data

Slice and Dice

In this example we are loading a tsv data set and using the load_dataset with csv method and passing in the \t delimiter to load the data

Filter Dataset (example remove None columns)

Refer to code for the example

Creating a new column

Creating a new column named review_length, which contains the lenght of each review.

Refer the module 5 code for an example

Sorting the dataset

Dataset have a method sort which can sort based on the column, see the code for example

The `map()` method’s superpowers

The code contains the example where we use batched=True to speed up the processing.

map function also has num_proc argument to run the computation parrelly

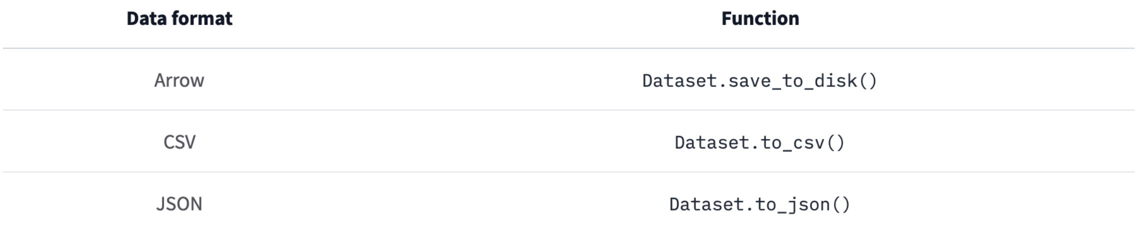
Dataset can be converted to a pandas df using:

Dataset.set_format('pandas')

Under the hood, `Dataset.set_format()` changes the return format for the dataset's `__getitem__()` dunder method. This means that when we want to create a new object like `train_df` from a `Dataset` in the `"pandas"` format, we need to slice the whole dataset to obtain a `pandas.DataFrame`. You can verify for yourself that the type of `drug_dataset["train"]` is `Dataset`, irrespective of the output format.

Creating a validation set

Saving a dataset



Data format	Function
Arrow	`Dataset.save_to_disk()`
CSV	`Dataset.to_csv()`
JSON	`Dataset.to_json()`