Load text data

This guide shows you how to load text datasets. To learn how to load any type of dataset, take a look at the general loading guide.

Text files are one of the most common file types for storing a dataset. By default, Datasets samples a text file line by line to build the dataset.

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

# Load from a directory
>>> dataset = load_dataset("text", data_dir="path/to/text/dataset")
```

To sample a text file by paragraph or even an entire document, use the `sample_by` parameter:

```
# Sample by paragraph
>>> dataset = load_dataset("text", data_files={"train": "my_train_file.txt", "test": "my_test_file

# Sample by document
>>> dataset = load_dataset("text", data_files={"train": "my_train_file.txt", "test": "my_test_file
```

You can also use grep patterns to load specific files:

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

To load remote text files via HTTP, pass the URLs instead:

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

To load XML data you can use the "xml" loader, which is equivalent to "text" with sample_by="document":

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

# Load from a directory
>>> dataset = load_dataset("xml", data_dir="path/to/xml/dataset")
```