Big data? 🤗 Datasets to the rescue!

File Edit View Insert Runtime Tools Help

+ Code + Text Copy to Drive

Big data? 🤗 Datasets to the rescue!

Install the Transformers, Datasets, and Evaluate libraries to run this notebook.

```
[ ] !pip install datasets evaluate transformers[sentencepiece]
```

```
[ ] !pip install zstandard
```

```
[ ] from datasets import load_dataset

    # This takes a few minutes to run, so go grab a tea or coffee while you wait :)
    data_files = "https://the-eye.eu/public/AI/pile_preliminary_components/PUBMED_title_abstracts_2019_baseline.jsonl.zst"
    pubmed_dataset = load_dataset("json", data_files=data_files, split="train")
    pubmed_dataset
```

```
Dataset({
    features: ['meta', 'text'],
    num_rows: 15518009
})
```

```
[ ] pubmed_dataset[0]
```

```
{'meta': {'pmid': 11409574, 'language': 'eng'},
 'text': 'Epidemiology of hypoxaemia in children with acute lower respiratory infection.\nTo determine the prevalence of hypoxa
lower respiratory infections (ALRI), the risk factors for hypoxaemia in children under 5 years of age with ALRI, and the associ
children of the same age ...'}
```

```
[ ] !pip install psutil
```

```
[ ] import psutil

    # Process.memory_info is expressed in bytes, so convert to megabytes
    print(f"RAM used: {psutil.Process().memory_info().rss / (1024 * 1024):.2f} MB")
```

```
RAM used: 5678.33 MB
```

```
[ ] print(f"Number of files in dataset : {pubmed_dataset.dataset_size}")
    size_gb = pubmed_dataset.dataset_size / (1024**3)
    print(f"Dataset size (cache file) : {size_gb:.2f} GB")
```

```
Number of files in dataset : 20979437051
Dataset size (cache file) : 19.54 GB
```

```
[ ] import timeit

    code_snippet = """batch_size = 1000

    for idx in range(0, len(pubmed_dataset), batch_size):
        _ = pubmed_dataset[idx:idx + batch_size]
    """

    time = timeit.timeit(stmt=code_snippet, number=1, globals=globals())
    print(
        f"Iterated over {len(pubmed_dataset)} examples (about {size_gb:.1f} GB) in "
        f"{time:.1f}s, i.e. {size_gb/time:.3f} GB/s"
    )
```

```
'Iterated over 15518009 examples (about 19.5 GB) in 64.2s, i.e. 0.304 GB/s'
```

```
[ ] pubmed_dataset_streamed = load_dataset(
        "json", data_files=data_files, split="train", streaming=True
    )
```

```
[ ] next(iter(pubmed_dataset_streamed))
```

```
{'meta': {'pmid': 11409574, 'language': 'eng'},
 'text': 'Epidemiology of hypoxaemia in children with acute lower respiratory infection.\nTo determine the prevalence of hypoxa
lower respiratory infections (ALRI), the risk factors for hypoxaemia in children under 5 years of age with ALRI, and the associ
children of the same age ...'}
```

```
[ ]  from transformers import AutoTokenizer

     tokenizer = AutoTokenizer.from_pretrained("distilbert-base-uncased")
     tokenized_dataset = pubmed_dataset_streamed.map(lambda x: tokenizer(x["text"]))
     next(iter(tokenized_dataset))
```

```
{'input_ids': [101, 4958, 5178, 4328, 6779, ...], 'attention_mask': [1, 1, 1, 1, 1, ...]}
```

```
[ ]  shuffled_dataset = pubmed_dataset_streamed.shuffle(buffer_size=10_000, seed=42)
     next(iter(shuffled_dataset))
```

```
{'meta': {'pmid': 11410799, 'language': 'eng'},
 'text': 'Randomized study of dose or schedule modification of granulocyte colony-stimulating factor in platinum-based chemothe
```

```
[ ]  dataset_head = pubmed_dataset_streamed.take(5)
     list(dataset_head)
```

```
[{'meta': {'pmid': 11409574, 'language': 'eng'},
  'text': 'Epidemiology of hypoxaemia in children with acute lower respiratory infection ...'},
 {'meta': {'pmid': 11409575, 'language': 'eng'},
  'text': 'Clinical signs of hypoxaemia in children with acute lower respiratory infection: indicators of oxygen therapy ...'},
 {'meta': {'pmid': 11409576, 'language': 'eng'},
  'text': "Hypoxaemia in children with severe pneumonia in Papua New Guinea ..."},
 {'meta': {'pmid': 11409577, 'language': 'eng'},
  'text': 'Oxygen concentrators and cylinders ...'},
 {'meta': {'pmid': 11409578, 'language': 'eng'},
  'text': 'Oxygen supply in rural africa: a personal experience ...'}]
```

```
[ ]  # Skip the first 1,000 examples and include the rest in the training set
     train_dataset = shuffled_dataset.skip(1000)
     # Take the first 1,000 examples for the validation set
     validation_dataset = shuffled_dataset.take(1000)
```

```
[ ]  law_dataset_streamed = load_dataset(
         "json",
         data_files="https://the-eye.eu/public/AI/pile_preliminary_components/FreeLaw_Opinions.jsonl.zst",
         split="train",
         streaming=True,
     )
     next(iter(law_dataset_streamed))
```

```
{'meta': {'case_ID': '110921.json',
  'case_jurisdiction': 'scotus.tar.gz',
  'date_created': '2010-04-28T17:12:49Z'},
 'text': '\n461 U.S. 238 (1983)\nOLIM ET AL.\nv.\nWAKINEKONA\nNo. 81-1581.\nSupreme Court of United States.\nArgued January 19,
UNITED STATES COURT OF APPEALS FOR THE NINTH CIRCUIT\n*239 Michael A. Lilly, First Deputy Attorney General of Hawaii, argued th
H. Dannenberg, Deputy Attorney General...'}
```

[ ]
```
from itertools import islice
from datasets import interleave_datasets

combined_dataset = interleave_datasets([pubmed_dataset_streamed, law_dataset_streamed])
list(islice(combined_dataset, 2))
```

```
[{'meta': {'pmid': 11409574, 'language': 'eng'},
  'text': 'Epidemiology of hypoxaemia in children with acute lower respiratory infection ...'},
 {'meta': {'case_ID': '110921.json',
   'case_jurisdiction': 'scotus.tar.gz',
   'date_created': '2010-04-28T17:12:49Z'},
  'text': '\n461 U.S. 238 (1983)\nOLIM ET AL.\nv.\nWAKINEKONA\nNo. 81-1581.\nSupreme Court of United States.\nArgued January 19
UNITED STATES COURT OF APPEALS FOR THE NINTH CIRCUIT\n*239 Michael A. Lilly, First Deputy Attorney General of Hawaii, argued th
H. Dannenberg, Deputy Attorney General...'}]
```

[ ]
```
base_url = "https://the-eye.eu/public/AI/pile/"
data_files = {
    "train": [base_url + "train/" + f"{idx:02d}.jsonl.zst" for idx in range(30)],
    "validation": base_url + "val.jsonl.zst",
    "test": base_url + "test.jsonl.zst",
}
pile_dataset = load_dataset("json", data_files=data_files, streaming=True)
next(iter(pile_dataset["train"]))
```

```
{'meta': {'pile_set_name': 'Pile-CC'},
 'text': 'It is done, and submitted. You can play “Survival of the Tastiest” on Android, and on the web...'}
```