06_BERT

March 12, 2024

```
[13]: !pip  install transformers tokenizers datasets accelerate evaluate
```

```
Requirement already satisfied: transformers in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(4.28.0)
Requirement already satisfied: tokenizers in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(0.13.3)
Requirement already satisfied: datasets in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(2.12.0)
Collecting accelerate
  Using cached accelerate-0.27.2-py3-none-any.whl.metadata (18 kB)
Requirement already satisfied: evaluate in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(0.4.1)
Requirement already satisfied: filelock in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from transformers) (3.13.1)
Requirement already satisfied: huggingface-hub<1.0,>=0.11.0 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from transformers) (0.20.3)
Requirement already satisfied: numpy>=1.17 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from transformers) (1.26.4)
Requirement already satisfied: packaging>=20.0 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from transformers) (23.2)
Requirement already satisfied: pyyaml>=5.1 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from transformers) (6.0.1)
Requirement already satisfied: regex!=2019.12.17 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from transformers) (2023.12.25)
Requirement already satisfied: requests in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from transformers) (2.31.0)
Requirement already satisfied: tqdm>=4.27 in
```

```
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from transformers) (4.66.2)
Requirement already satisfied: pyarrow>=8.0.0 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from datasets) (15.0.0)
Requirement already satisfied: dill<0.3.7,>=0.3.0 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from datasets) (0.3.6)
Requirement already satisfied: pandas in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from datasets) (2.2.0)
Requirement already satisfied: xxhash in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from datasets) (3.4.1)
Requirement already satisfied: multiprocess in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from datasets) (0.70.14)
Requirement already satisfied: fsspec>=2021.11.1 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from fsspec[http]>=2021.11.1->datasets) (2023.10.0)
Requirement already satisfied: aiohttp in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from datasets) (3.9.3)
Requirement already satisfied: responses<0.19 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from datasets) (0.18.0)
Requirement already satisfied: psutil in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from accelerate) (5.9.8)
Requirement already satisfied: torch>=1.10.0 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from accelerate) (2.2.0)
Requirement already satisfied: safetensors>=0.3.1 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from accelerate) (0.4.2)
Requirement already satisfied: aiosignal>=1.1.2 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from aiohttp->datasets) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from aiohttp->datasets) (23.2.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from aiohttp->datasets) (1.4.1)
Requirement already satisfied: multidict<7.0,>=4.5 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from aiohttp->datasets) (6.0.5)
Requirement already satisfied: yarl<2.0,>=1.0 in
```

```
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from aiohttp->datasets) (1.9.4)
Requirement already satisfied: async-timeout<5.0,>=4.0 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from aiohttp->datasets) (4.0.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from huggingface-hub<1.0,>=0.11.0->transformers) (4.9.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from requests->transformers) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from requests->transformers) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from requests->transformers) (2.2.0)
Requirement already satisfied: certifi>=2017.4.17 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from requests->transformers) (2024.2.2)
Requirement already satisfied: sympy in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from torch>=1.10.0->accelerate) (1.12)
Requirement already satisfied: networkx in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from torch>=1.10.0->accelerate) (3.2.1)
Requirement already satisfied: jinja2 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from torch>=1.10.0->accelerate) (3.1.3)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.1.105 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from torch>=1.10.0->accelerate) (12.1.105)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.1.105 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from torch>=1.10.0->accelerate) (12.1.105)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.1.105 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from torch>=1.10.0->accelerate) (12.1.105)
Requirement already satisfied: nvidia-cudnn-cu12==8.9.2.26 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from torch>=1.10.0->accelerate) (8.9.2.26)
Requirement already satisfied: nvidia-cublas-cu12==12.1.3.1 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from torch>=1.10.0->accelerate) (12.1.3.1)
Requirement already satisfied: nvidia-cufft-cu12==11.0.2.54 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from torch>=1.10.0->accelerate) (11.0.2.54)
Requirement already satisfied: nvidia-curand-cu12==10.3.2.106 in
```

```
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from torch>=1.10.0->accelerate) (10.3.2.106)
Requirement already satisfied: nvidia-cusolver-cu12==11.4.5.107 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from torch>=1.10.0->accelerate) (11.4.5.107)
Requirement already satisfied: nvidia-cusparse-cu12==12.1.0.106 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from torch>=1.10.0->accelerate) (12.1.0.106)
Requirement already satisfied: nvidia-nccl-cu12==2.19.3 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from torch>=1.10.0->accelerate) (2.19.3)
Requirement already satisfied: nvidia-nvtx-cu12==12.1.105 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from torch>=1.10.0->accelerate) (12.1.105)
Requirement already satisfied: triton==2.2.0 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from torch>=1.10.0->accelerate) (2.2.0)
Requirement already satisfied: nvidia-nvjitlink-cu12 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from nvidia-cusolver-cu12==11.4.5.107->torch>=1.10.0->accelerate) (12.3.101)
Requirement already satisfied: python-dateutil>=2.8.2 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from pandas->datasets) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from pandas->datasets) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from pandas->datasets) (2024.1)
Requirement already satisfied: six>=1.5 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from python-dateutil>=2.8.2->pandas->datasets) (1.16.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from jinja2->torch>=1.10.0->accelerate) (2.1.5)
Requirement already satisfied: mpmath>=0.19 in
/home/solaris/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages
(from sympy->torch>=1.10.0->accelerate) (1.3.0)
Using cached accelerate-0.27.2-py3-none-any.whl (279 kB)
Installing collected packages: accelerate
Successfully installed accelerate-0.27.2
```

[14]:

```python
import evaluate
import numpy as np
import torch
from transformers import AutoTokenizer, DataCollatorWithPadding
from transformers import AutoModelForSequenceClassification
from transformers import Trainer
from transformers import TrainingArguments
from datasets import load_dataset
```

[15]:
```
raw_datsets = load_dataset("imdb")
```

```
Downloading and preparing dataset None/plain_text to file:///home/solaris/.cache
/huggingface/datasets/parquet/plain_text-
745310791ff4d097/0.0.0/2a3b91fbd88a2c90d1dbbb32b460cf621d31bd5b05b934492fdef7d8d
6f236ec…

Downloading data files:   0%|          | 0/2 [00:00<?, ?it/s]

Extracting data files:   0%|          | 0/2 [00:00<?, ?it/s]

Generating train split:   0%|          | 0/25000 [00:00<?, ? examples/s]

Generating test split:   0%|          | 0/25000 [00:00<?, ? examples/s]
```

```
---------------------------------------------------------------------------
ExpectedMoreSplits                        Traceback (most recent call last)
Cell In[15], line 1
----> 1 raw_datsets = load_dataset("imdb")

File ~/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages/datasets/
 ↪load.py:1797, in load_dataset(path, name, data_dir, data_files, split,␣
 ↪cache_dir, features, download_config, download_mode, verification_mode,␣
 ↪ignore_verifications, keep_in_memory, save_infos, revision, use_auth_token,␣
 ↪task, streaming, num_proc, storage_options, **config_kwargs)
   1794 try_from_hf_gcs = path not in _PACKAGED_DATASETS_MODULES
   1796 # Download and prepare data
-> 1797 builder_instance.download_and_prepare(
   1798     download_config=download_config,
   1799     download_mode=download_mode,
   1800     verification_mode=verification_mode,
   1801     try_from_hf_gcs=try_from_hf_gcs,
   1802     num_proc=num_proc,
   1803     storage_options=storage_options,
   1804 )
   1806 # Build dataset for splits
   1807 keep_in_memory = (
   1808     keep_in_memory if keep_in_memory is not None else␣
 ↪is_small_dataset(builder_instance.info.dataset_size)
   1809 )

File ~/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages/datasets/
 ↪builder.py:890, in DatasetBuilder.download_and_prepare(self, output_dir,␣
 ↪download_config, download_mode, verification_mode, ignore_verifications,␣
 ↪try_from_hf_gcs, dl_manager, base_path, use_auth_token, file_format,␣
 ↪max_shard_size, num_proc, storage_options, **download_and_prepare_kwargs)
    888     if num_proc is not None:
```

```
    889         prepare_split_kwargs["num_proc"] = num_proc
--> 890     self._download_and_prepare(
    891         dl_manager=dl_manager,
    892         verification_mode=verification_mode,
    893         **prepare_split_kwargs,
    894         **download_and_prepare_kwargs,
    895     )
    896 # Sync info
    897 self.info.dataset_size = sum(split.num_bytes for split in self.info.splits.values())

File ~/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages/datasets/builder.py:1003, in DatasetBuilder._download_and_prepare(self, dl_manager, verification_mode, **prepare_split_kwargs)
   1000     dl_manager.manage_extracted_files()
   1002 if verification_mode == VerificationMode.BASIC_CHECKS or verification_mode == VerificationMode.ALL_CHECKS:
-> 1003     verify_splits(self.info.splits, split_dict)
   1005 # Update the info object with the splits.
   1006 self.info.splits = split_dict

File ~/miniconda3/envs/deep_learning_nlp/lib/python3.10/site-packages/datasets/utils/info_utils.py:91, in verify_splits(expected_splits, recorded_splits)
     89     return
     90 if len(set(expected_splits) - set(recorded_splits)) > 0:
---> 91     raise ExpectedMoreSplits(str(set(expected_splits) - set(recorded_splits)))
     92 if len(set(recorded_splits) - set(expected_splits)) > 0:
     93     raise UnexpectedSplits(str(set(recorded_splits) - set(expected_splits)))

ExpectedMoreSplits: {'unsupervised'}
```