Datasets: legacy-datasets/**wikipedia**

♡ like 591

Follow Legacy Datasets 31

Tasks: Text Generation Fill-Mask

Sub-tasks: language-modeling masked-language-modeling

Languages: Afar Abkhaz Achinese + 291 Size: n<1K

License: cc-by-sa-3.0 gfdl

Dataset card Files Community 22

Downloads last month 29,134

Homepage:
dumps.wikimedia.org

Models trained or fine-tuned on `legacy-datasets/wikipedia`

`google-bert/bert-base-uncased`
Fill-Mask • Updated Feb 19, 2024 • 66.8M • ♡ 2.23k

`FacebookAI/roberta-large`
Fill-Mask • Updated Feb 19, 2024 • 18.6M • ♡ 210

`distilbert/distilbert-base-uncased`
Fill-Mask • Updated May 6, 2024 • 12M • ♡ 666

`FacebookAI/roberta-base`
Fill-Mask • Updated Feb 19, 2024 • 9.14M • ♡ 485

Browse 1756 models trained on this dataset

Spaces using `legacy-datasets/wikipedia` 22

zjunlp/EasyEdit WordLift/entity-linking kikuepi/dify xinzhang1219/test240527 Severian/dify + 17 Spaces

The Dataset Viewer has been disabled on this dataset.

Dataset Card for Wikipedia

Dataset Summary

Wikipedia dataset containing cleaned articles of all languages. The datasets are built from the Wikipedia dump (https://dumps.wikimedia.org/) with one split per language. Each example contains the content of one full Wikipedia article with cleaning to strip markdown and unwanted sections (references, etc.).

The articles are parsed using the `mwparserfromhell` tool, which can be installed with:

```
pip install mwparserfromhell
```

Then, you can load any subset of Wikipedia per language and per date this way:

```
from datasets import load_dataset

load_dataset("wikipedia", language="sw", date="20220120")
```

> You can specify `num_proc=` in `load_dataset` to generate the dataset in parallel.

You can find the full list of languages and dates here.

Some subsets of Wikipedia have already been processed by HuggingFace, and you can load them just with:

```
from datasets import load_dataset

load_dataset("wikipedia", "20220301.en")
```

The list of pre-processed subsets is:

- "20220301.de"
- "20220301.en"
- "20220301.fr"
- "20220301.frr"
- "20220301.it"
- "20220301.simple"

Supported Tasks and Leaderboards

The dataset is generally used for Language Modeling.

Languages

You can find the list of languages here.

Dataset Structure

Data Instances

An example looks as follows:

```
{'id': '1',
 'url': 'https://simple.wikipedia.org/wiki/April',
 'title': 'April',
 'text': 'April is the fourth month...'
}
```

Some subsets of Wikipedia have already been processed by HuggingFace, as you can see below:

20220301.de

- **Size of downloaded dataset files:** 5.34 GB
- **Size of the generated dataset:** 8.91 GB
- **Total amount of disk used:** 14.25 GB

20220301.en

- **Size of downloaded dataset files:** 11.69 GB
- **Size of the generated dataset:** 20.28 GB
- **Total amount of disk used:** 31.96 GB

20220301.fr

- **Size of downloaded dataset files:** 4.22 GB
- **Size of the generated dataset:** 7.38 GB
- **Total amount of disk used:** 11.60 GB

20220301.frr

- **Size of downloaded dataset files:** 4.53 MB
- **Size of the generated dataset:** 9.13 MB
- **Total amount of disk used:** 13.66 MB

20220301.it

- **Size of downloaded dataset files:** 2.71 GB
- **Size of the generated dataset:** 4.54 GB

- **Total amount of disk used:** 7.25 GB

20220301.simple

- **Size of downloaded dataset files:** 133.89 MB
- **Size of the generated dataset:** 235.07 MB
- **Total amount of disk used:** 368.96 MB

Data Fields

The data fields are the same among all configurations:

- `id` (`str`): ID of the article.
- `url` (`str`): URL of the article.
- `title` (`str`): Title of the article.
- `text` (`str`): Text content of the article.

Data Splits

Here are the number of examples for several configurations:

name	train
20220301.de	2665357
20220301.en	6458670
20220301.fr	2402095
20220301.frr	15199
20220301.it	1743035
20220301.simple	205328

Dataset Creation

Curation Rationale

More Information Needed

Source Data

Initial Data Collection and Normalization

More Information Needed

Who are the source language producers?

More Information Needed

Annotations

Annotation process

More Information Needed

Who are the annotators?

More Information Needed

Personal and Sensitive Information

More Information Needed

Considerations for Using the Data

Social Impact of Dataset

More Information Needed

Discussion of Biases

More Information Needed

Other Known Limitations

More Information Needed

Additional Information

Dataset Curators

More Information Needed

Licensing Information

Most of Wikipedia's text and many of its images are co-licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported License (CC BY-SA) and the GNU Free Documentation License (GFDL) (unversioned, with no invariant sections, front-cover texts, or back-cover texts).

Some text has been imported only under CC BY-SA and CC BY-SA-compatible license and cannot be reused under GFDL; such text will be identified on the page footer, in the page history, or on the discussion page of the article that utilizes the text.

Citation Information

```
@ONLINE{wikidump,
    author = "Wikimedia Foundation",
    title  = "Wikimedia Downloads",
    url    = "https://dumps.wikimedia.org"
}
```

Contributions

Thanks to @lewtun, @mariamabarham, @thomwolf, @lhoestq, @patrickvonplaten for adding this dataset.