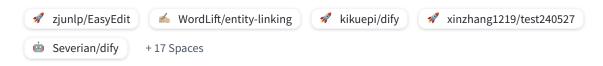


Browse 1756 models trained on this dataset

■ Spaces using legacy-datasets/wikipedia 22



The Dataset Viewer has been disabled on this dataset.

## Dataset Card for Wikipedia

### Dataset Summary

 $\equiv$ 

Wikipedia dataset containing cleaned articles of all languages. The datasets are built from the Wikipedia dump (<a href="https://dumps.wikimedia.org/">https://dumps.wikimedia.org/</a>) 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 <u>here</u>.

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.

# 

You can find the list of languages <u>here</u>.

### Dataset Structure

### 

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

### O 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.

### O 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

More Information Needed
⊘ Source Data
⊘ Initial Data Collection and Normalization
More Information Needed
More Information Needed
<b>⊘</b> 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
Ø 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 <u>Creative</u> <u>Commons Attribution-ShareAlike 3.0 Unported License</u> (CC BY-SA) and the <u>GNU Free</u> <u>Documentation License</u> (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 <u>@lewtun</u>, <u>@mariamabarham</u>, <u>@thomwolf</u>, <u>@lhoestq</u>, <u>@patrickvonplaten</u> for adding this dataset.

# ■ System theme

# Company

TOS

Privacy

About

Jobs

# Website

Models

Datasets

Spaces

Pricing

Docs

