

TFDS now supports the [Croissant](https://mlcommons.org/croissant) format (<https://mlcommons.org/croissant>)! Read the [documentation](https://www.tensorflow.org/datasets/format_specific_dataset_builders#croissantbuilder) (https://www.tensorflow.org/datasets/format_specific_dataset_builders#croissantbuilder) to know more.

user_libri_text

- **Description:**

UserLibri is a dataset containing paired audio-transcripts and additional text only data for each of 107 users. It is a reformatting of the LibriSpeech dataset found at <http://www.openslr.org/12>, (<http://www.openslr.org/12>,) reorganizing the data into users with an average of 52 LibriSpeech utterances and about 6,700 text example sentences per user. The UserLibriAudio class provides access to the audio-transcript pairs. See UserLibriText for the additional text data.

- **Homepage:** <https://www.kaggle.com/datasets/google/userlibri>
(<https://www.kaggle.com/datasets/google/userlibri>)
- **Source code:** `tfds.text.userlibri_lm_data.UserLibriText`
(https://github.com/tensorflow/datasets/tree/master/tensorflow_datasets/text/userlibri_lm_data/userlibri_lm_data.py)
- **Versions:**
 - **1.0.0** (default): No release notes.
- **Download size:** Unknown size
- **Dataset size:** 86.86 MiB
- **Auto-cached** ([documentation](https://www.tensorflow.org/datasets/performances#auto-caching) (<https://www.tensorflow.org/datasets/performances#auto-caching>)): Yes
- **Splits:**

Split	Examples
'10136'	38,496
'1041'	970
'10540'	3,283

Split	Examples
' 108 '	5,864
' 11 '	1,348
' 11667 '	3,312
' 1184 '	22,062
' 12176 '	1,467
' 12434 '	2,796
' 12544 '	4,080
' 13110 '	2,634
' 13158 '	3,440
' 13441 '	4,145
' 135 '	37,263
' 1353 '	4,889
' 1399 '	18,914
' 14420 '	6,950
' 14566 '	3,810
' 1477 '	2,526
' 14958 '	1,495
' 15263 '	21,085
' 15265 '	7,647
' 1549 '	5,439
' 1572 '	2,882
' 1597 '	3,586
' 1608 '	3,605
' 16127 '	3,588
' 16653 '	7,600

Split	Examples
'18096'	2,384
'1827'	4,806
'19019'	3,248
'19215'	13,542
'19717'	3,762
'1989'	1,105
'1998'	8,923
'20019'	966
'2002'	239
'20212'	3,363
'209'	2,090
'21297'	4,165
'22002'	4,044
'2300'	22,201
'24'	3,537
'24585'	1,789
'24811'	2,399
'2488'	8,239
'2529'	3,934
'26177'	3,598
'26379'	379
'2681'	8,872
'27067'	3,149
'27090'	3,217
'2770'	3,750

Split	Examples
' 2787 '	4,603
' 28700 '	5,547
' 28725 '	3,899
' 28952 '	2,909
' 2981 '	54,305
' 3076 '	7,124
' 30905 '	2,140
' 3178 '	8,454
' 33 '	3,569
' 33800 '	5,145
' 3436 '	5,899
' 3440 '	5,087
' 3441 '	6,042
' 36508 '	521
' 3748 '	4,767
' 38675 '	2,696
' 38804 '	5,653
' 39159 '	2,729
' 4028 '	9,633
' 40359 '	7,821
' 41326 '	6,181
' 4217 '	6,003
' 4276 '	10,461
' 434 '	4,319
' 4602 '	4,421

Split	Examples
'507'	9,093
'540'	5,452
'5516'	4,963
'5630'	1,130
'574'	452
'5921'	6,040
'6328'	5,926
'6812'	5,839
'732'	22,971
'76'	6,454
'7891'	1,476
'8166'	3,190
'820'	11,054
'833'	3,638
'9189'	8,387
'94'	1,722
'940'	6,172
'9464'	1,695
'955'	3,051
'969'	7,799
'9983'	8,898

- **Feature structure:**

```
FeaturesDict({
    'book_id': Text(shape=(), dtype=string),
```

```
'text': Text(shape=(), dtype=string),
})
```

- **Feature documentation:**

Feature	Class	Shape	Dtype	Description
FeaturesDict				
book_id	Text		string	The book that this text was pulled from
text	Text		string	A sentence of text extracted from a book

- **Supervised keys** (See [as_supervised doc](#) (https://www.tensorflow.org/datasets/api_docs/python/tfds/load#args)): ('text', 'text')
- **Figure** ([tfds.show_examples](#) (https://www.tensorflow.org/datasets/api_docs/python/tfds/visualization/show_examples)): Not supported.
- **Examples** ([tfds.as_dataframe](#) (https://www.tensorflow.org/datasets/api_docs/python/tfds/as_dataframe)):

[Display examples...](#)

- **Citation:**

```
@inproceedings{breiner2022userlibri,
  title={UserLibri: A Dataset for ASR Personalization Using Only Text},
  author={Breiner, Theresa and Ramaswamy, Swaroop and Variani, Ehsan and Garg, Shefal},
  booktitle={Proc. Interspeech 2022},
  year={2022}
}
```

Except as otherwise noted, the content of this page is licensed under the [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/) (<https://creativecommons.org/licenses/by/4.0/>), and code samples are licensed under the [Apache 2.0 License](https://www.apache.org/licenses/LICENSE-2.0) (<https://www.apache.org/licenses/LICENSE-2.0>). For details, see the [Google Developers Site Policies](https://developers.google.com/site-policies) (<https://developers.google.com/site-policies>). Java is a registered trademark of Oracle and/or its affiliates.

Last updated 2022-12-06 UTC.