**Book corpus dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **# sentence** | **# words** | **# unique words** | **Average # words in each sentence** |
| **Book 1** | **40858231** | **440449278** | **496161** | **12.138** |
|
| **Book 2** | **34776361** | **364222279** | **382510** | **11.900** |
| **All** | **75634592** | **804671557** | **612194** | **12.019** |

**Link to get Book corpus dataset: https://storage.googleapis.com/huggingface-nlp/datasets/bookcorpus/bookcorpus.tar.bz2**