HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-docslite.de.tsv	7/5/2024	German (de)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
101,414,657	12,137,317,169			748.38 GB	

Top 10 domains

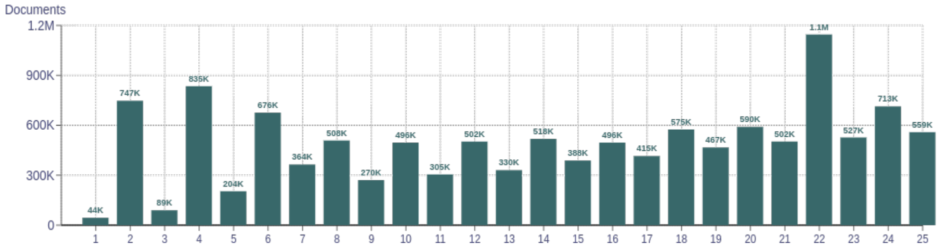
Domain	Docs	% of total
namendb.com	2.9M	2.87
blogspot.de	1.4M	1.37
blogspot.co.at	584K	0.58
diebuchsuche.com	580K	0.57
wordpress.com	569K	0.56
wikipedia.org	533K	0.53
blogspot.com	419K	0.41
hifi-forum.de	361K	0.36
blogspot.ch	341K	0.34
docplayer.org	323K	0.32

Top 10 TLDs

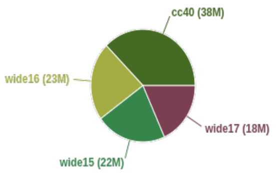
Domain	Docs	% of total
de	56M	54.99
com	21M	20.84
ch	5M	4.96
at	4.9M	4.84
org	3.1M	3.10
net	2.7M	2.62
eu	1.3M	1.28
info	1.3M	1.24
co.at	683K	0.67
club	284K	0.28

Documents size (in segments)

<= 25 segments **12.1%** (12M documents)
> 25 segments **87.9%** (89M documents)

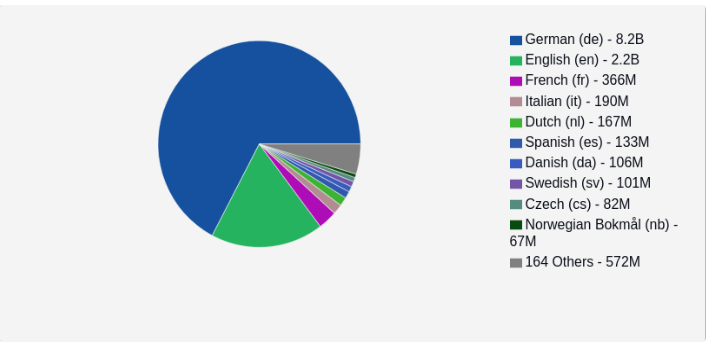


Documents by collection

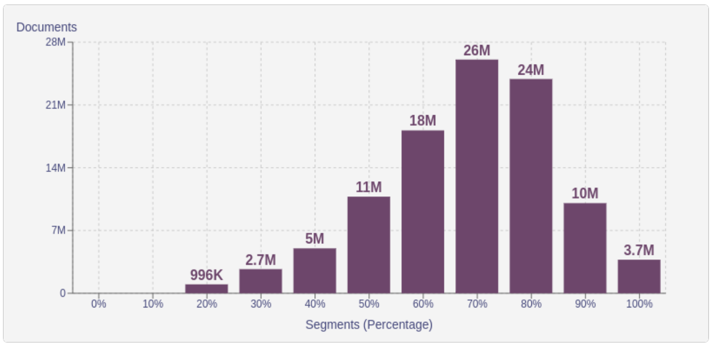


Language Distribution

Number of segments

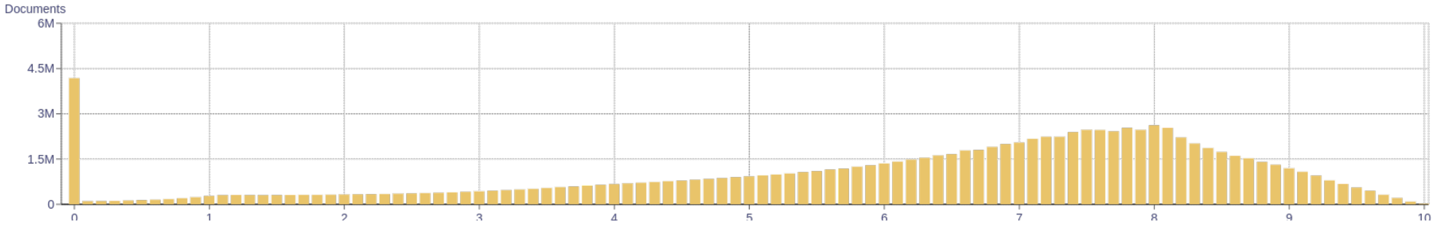


Percentage of segments in German (de) inside documents

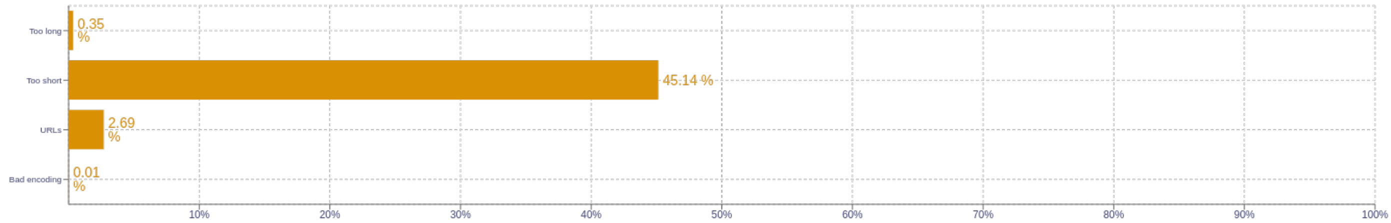


Distribution of documents by document score

score < 5 - **24.96%** (25M documents)
score >= 5 - **75.04%** (76M documents)



Segment noise distribution



About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, <ul>, <ol>, etc.) replaced by newlines.

Volumes - Tokens

Tokenized with https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md

Type-Token Ratio

Lexical variety computed as *number or types (uniques)/number of tokens*, after removing punctuation (https://www.sltinfo.com/wp-content/uploads/2014/01/type-token-ratio.pdf).

Document size (in segments)

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, <ul>, <ol>, etc.) replaced by newlines.

Language distribution

Language identified with FastSpell (https://github.com/mbanon/fastspell).

Distribution of segments by fluency score

Obtained with Monocleaner (https://github.com/bitextor/monocleaner).

Distribution of documents by average fluency score

Obtained with Monocleaner (https://github.com/bitextor/monocleaner).

Distribution of documents by document score

Obtained with Web Docs Scorer (https://github.com/pablop16n/web-docs-scorer/).

Segment length distribution by token

Tokenized with https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md

Segment noise distribution

Obtained with Bicleaner Hardrules (https://github.com/bitextor/bicleaner-hardrules/).

Frequent n-grams

Tokenized with https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md, after removing n-grams starting or ending in a stopword. Stopwords from https://github.com/hplt-project/data-analytics-tool/blob/main/scripts/resources/README.txt