

General overview

Corpus	Analytics date	Language
HPLT-v2-zho_Hant.tsv	9/30/2024	Chinese (zh)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
157,114,178	4,479,583,850			688.45 GB	282,496,310,558

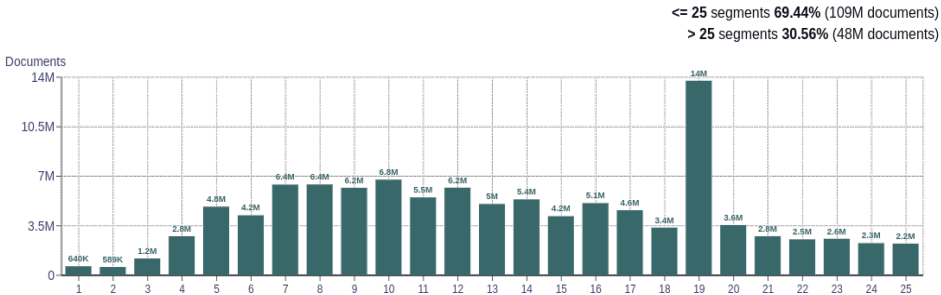
Top 10 domains

Domain	Docs	% of total
songtiankeji1.com	10M	6.34
pixnet.net	7.8M	4.97
b111.net	3.7M	2.34
blogspot.tw	1.6M	0.99
blogspot.com	1.1M	0.67
pscyxhd.com	849K	0.54
blogspot.hk	813K	0.52
zjggs.com	646K	0.41
eyny.com	607K	0.39
yahoo.com	593K	0.38

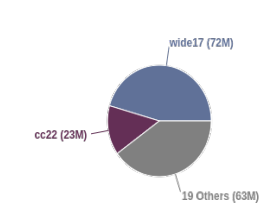
Top 10 TLDs

Domain	Docs	% of total
com	113M	71.83
net	16M	9.98
com.tw	7.1M	4.51
tw	5.1M	3.23
org	2.8M	1.79
hk	1.7M	1.05
cn	1.3M	0.80
com.hk	1M	0.66
click	756K	0.48
com.cn	732K	0.47

Documents size (in segments)

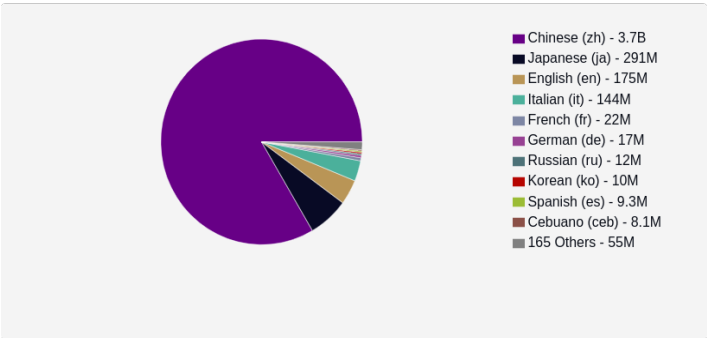


Documents by collection

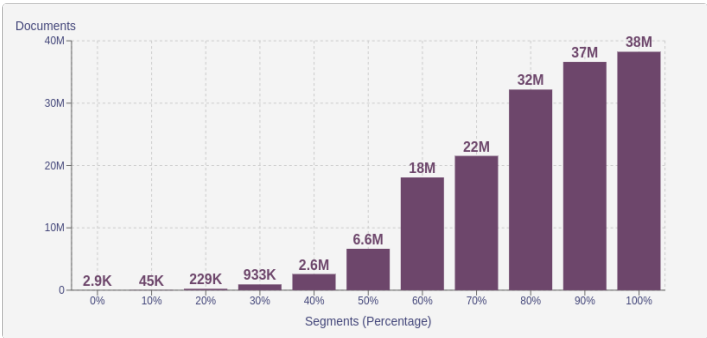


Language Distribution

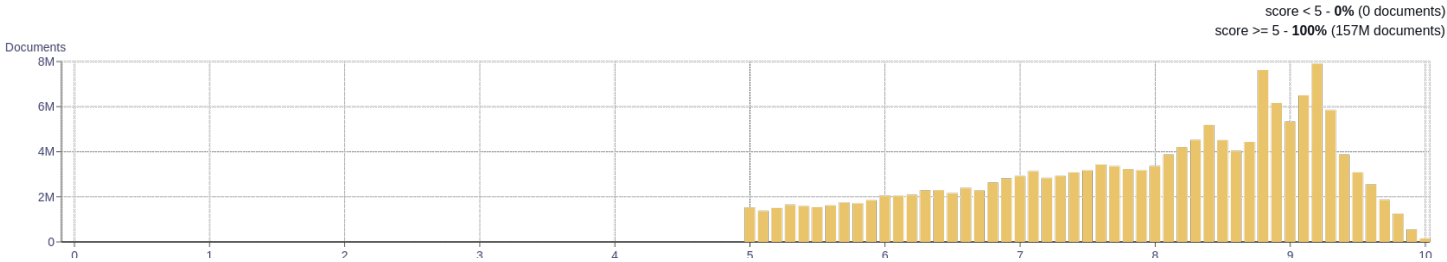
Number of segments



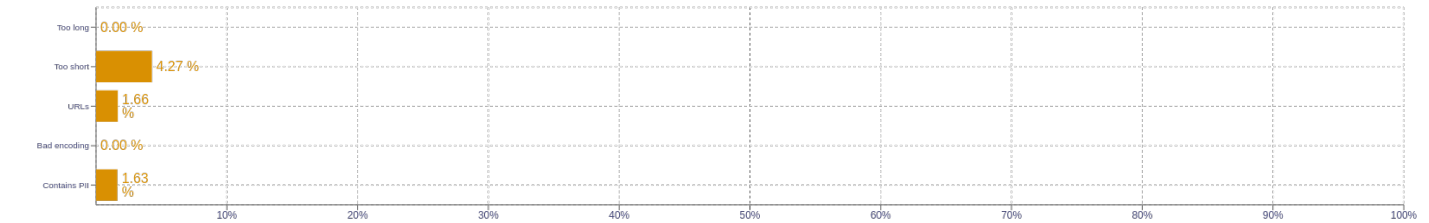
Percentage of segments in Chinese (zh) inside documents



Distribution of documents by document score



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 of 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>