

General overview

Corpus	Date	Language
HPLT-v2-zho_Hans.tsv	11/24/2024	Chinese (zh)

Volumes

Docs	Segments	Characters	Size
1,246,525,955	42,446,650,591	2,310,632,748,356	5.64 TB

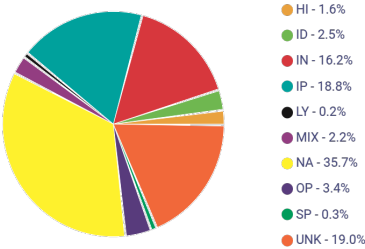
Top 10 domains

Domain	Docs	% of total
ganji.com	6.3M	0.50%
58.com	5.6M	0.45%
sina.com.cn	5.1M	0.41%
baixing.com	5.1M	0.41%
woaifexiang.net	4.8M	0.38%
ifeng.com	4.5M	0.36%
checheng123.com	4.3M	0.35%
520xs.com	4M	0.32%
163.com	3.8M	0.31%
shushu.com.cn	2.9M	0.23%

Top 10 TLDs

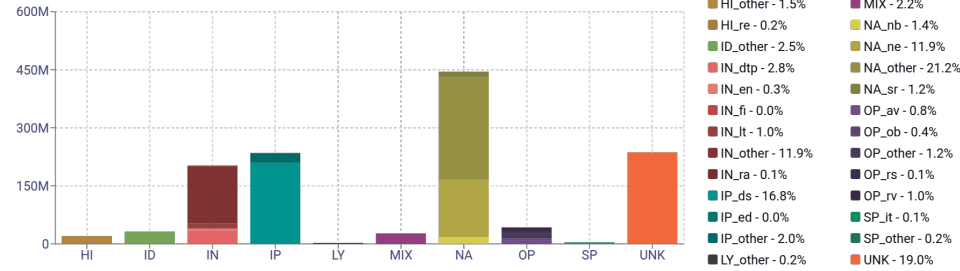
Domain	Docs	% of total
com	840M	67.37%
cn	154M	12.34%
net	72M	5.80%
com.cn	71M	5.68%
cc	24M	1.95%
org	23M	1.85%
gov.cn	11M	0.89%
top	5M	0.40%
org.cn	4.6M	0.37%
la	4.4M	0.36%

Register labels

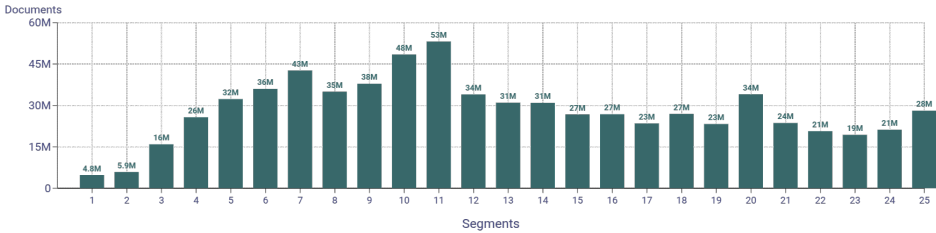


MT:7.0% | 87M Documents

Documents

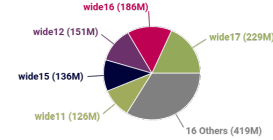


Documents size (in segments)



<= 25 segments 56.78% (708M documents)  
> 25 segments 43.22% (539M documents)

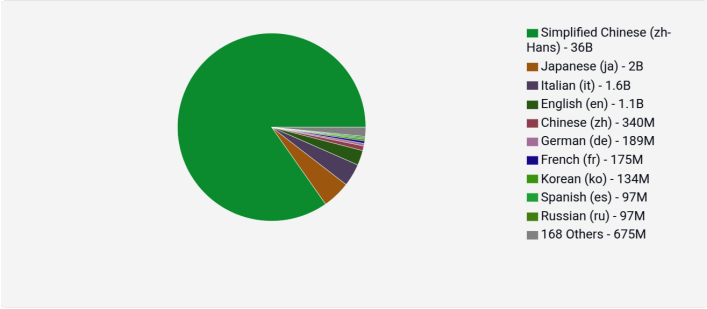
Documents by collection



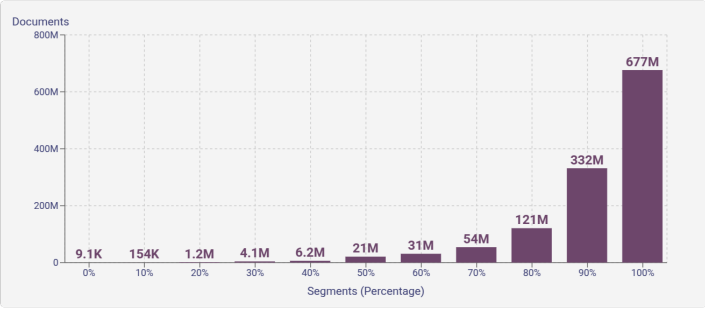
CC = 19.31%  
IA = 80.69%

Language Distribution

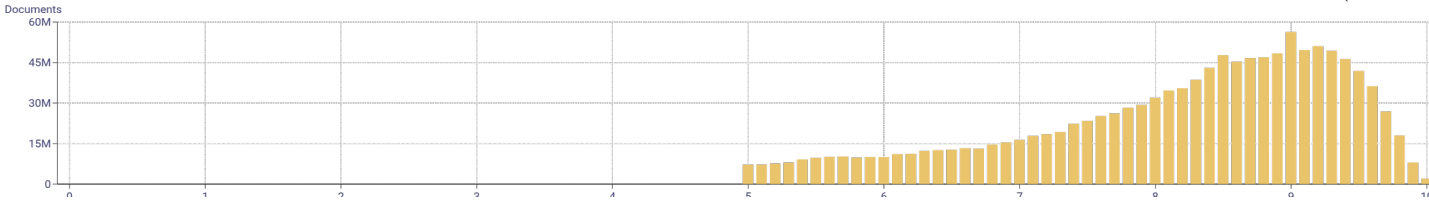
Number of segments in the Chinese (zh) corpus



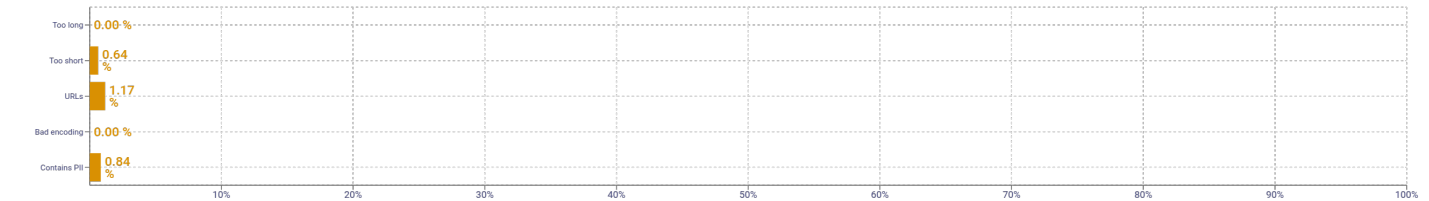
Percentage of segments in Chinese (zh) inside documents



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>

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	How-to or instructions	HI	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	Informational persuasion	IP	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	Opinion	OP
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	Informational description	IN	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				