

General overview

Corpus	Date	Language
hplt-v3-cmn_Hans	10/3/2025	Mandarin Chinese

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
578,000	16,238,063	14,388,982 (88.61 %)	11.39%	652M	1,108,021,352	2.88 GB

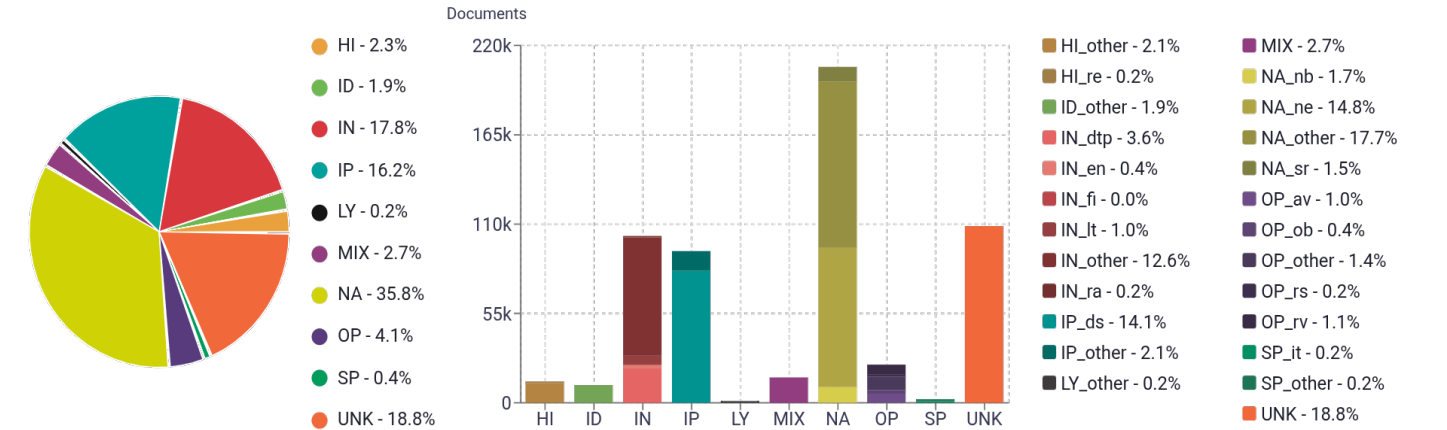
Top 10 domains

Domain	Docs	% of total
163.com	3K	0.51%
sina.com.cn	2.3K	0.39%
woaifenxiang.net	1.6K	0.28%
sohu.com	1.4K	0.25%
ifeng.com	1.4K	0.25%
58.com	1.4K	0.24%
checheng123.com	1.4K	0.24%
520xs.com	1.3K	0.22%
86zw.org	972	0.17%
kushubao.com	898	0.16%

Top 10 TLDs

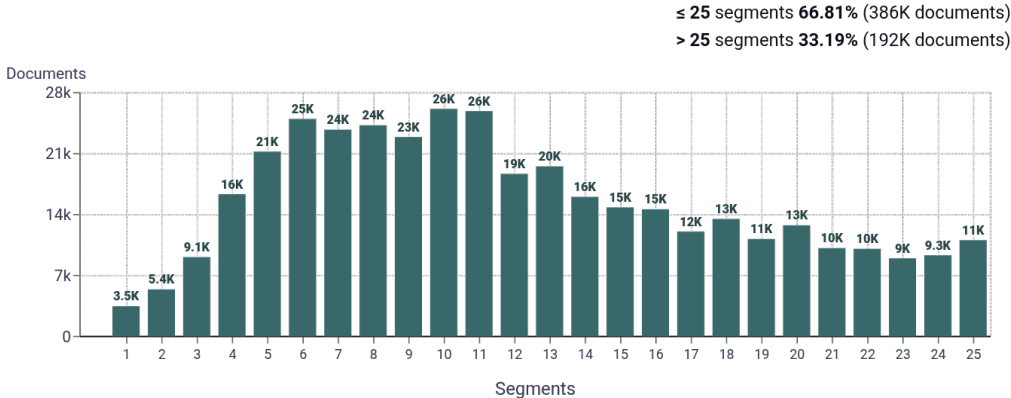
Domain	Docs	% of total
com	392K	67.81%
cn	70K	12.17%
net	34K	5.84%
com.cn	30K	5.13%
org	12K	2.00%
cc	8.8K	1.52%
gov.cn	5.7K	0.98%
edu.cn	2.8K	0.48%
org.cn	2.5K	0.43%
top	2.5K	0.43%

Register labels

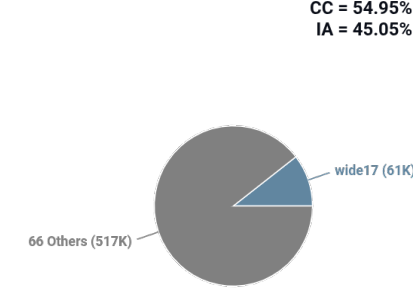


MT:5.9% | 34K Documents

Documents size (in segments) ⓘ

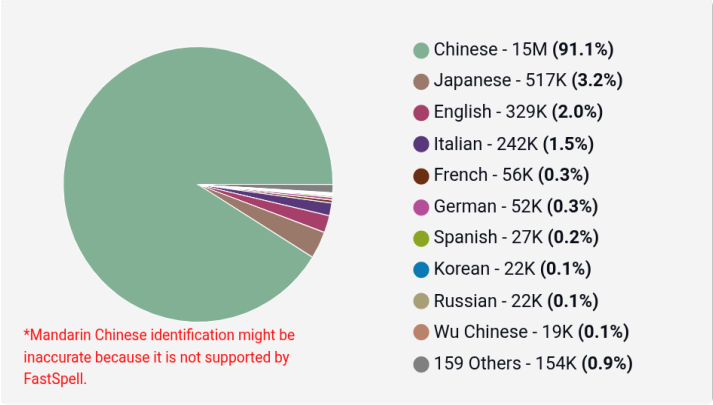


Document collections

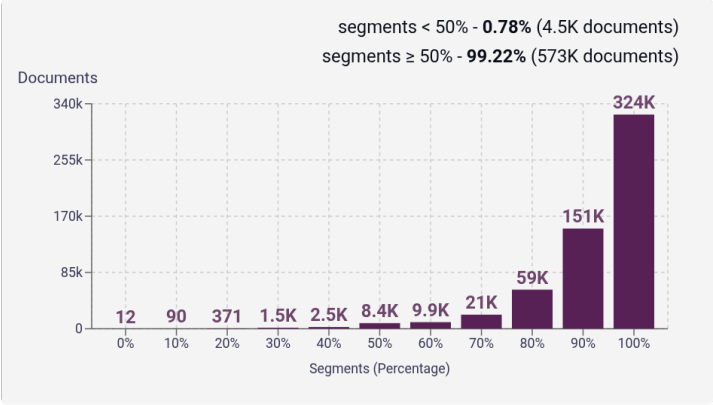


Language Distribution

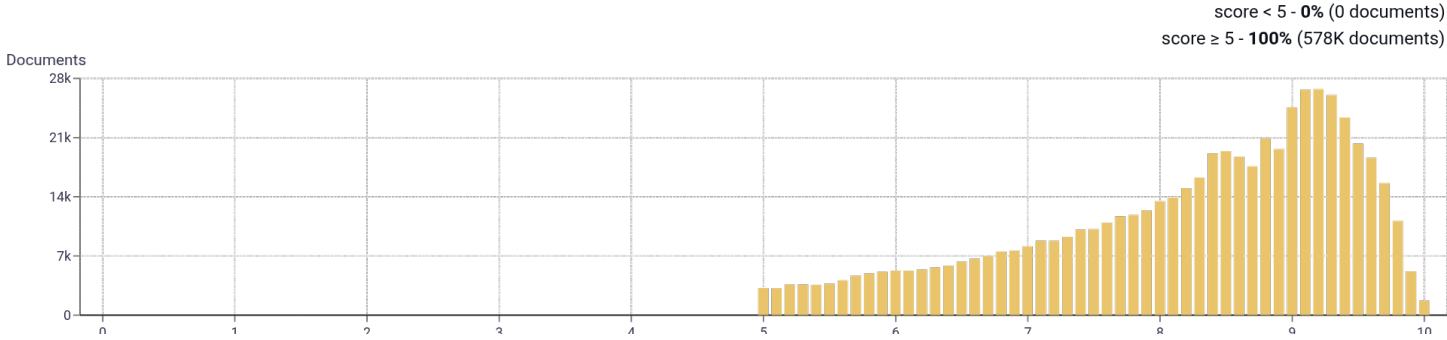
Number of segments in the Mandarin Chinese corpus



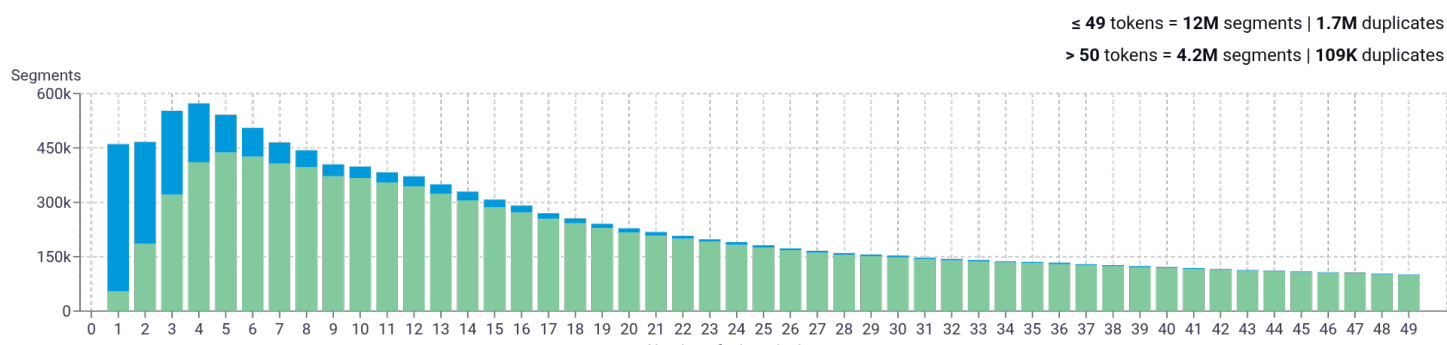
Percentage of segments in Mandarin Chinese inside documents



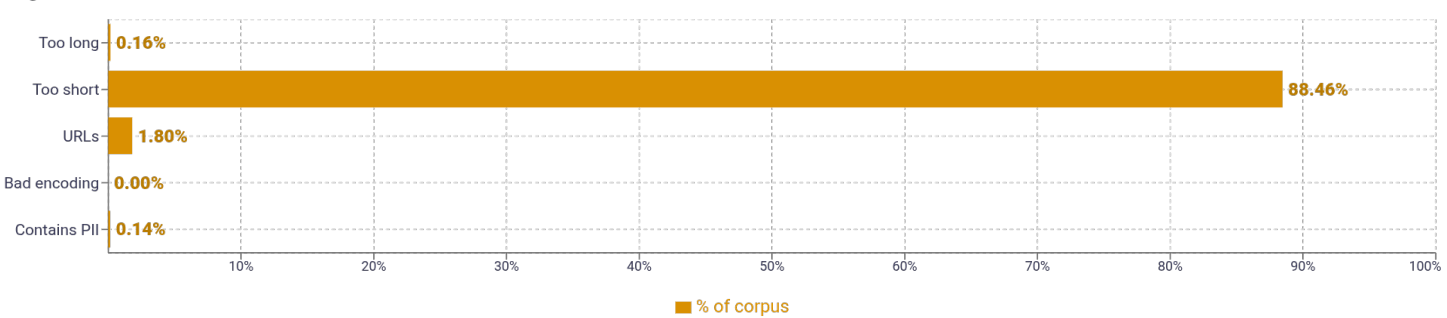
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	不   3,354,755 都   1,736,930 人   1,718,817 中   1,696,207 上   1,492,010	
2	不会   272,141 有限公司   205,617 最大   126,348 北京 pk   111,647 时时彩   110,672	
3	重庆时时彩   58,759 北京 pk 拾   25,068 股份有限公司   23,135 科技有限公司   21,637 会不会   18,195	
4	中国特色社会主义思想   6,547 时代中国特色社会主义   5,934 习近平新时代中国   5,878 平新时代中国特色   5,866 很大程度上   4,632	
5	习近平新时代中国特色   5,855 平新时代中国特色社会主义   5,835 时代中国特色社会主义思想   5,742 同志为核心的党中央   2,663 下载附件保存到相册   2,403	

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				