HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-cmn_Hans	10/3/2025	Chinese (cmn)

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

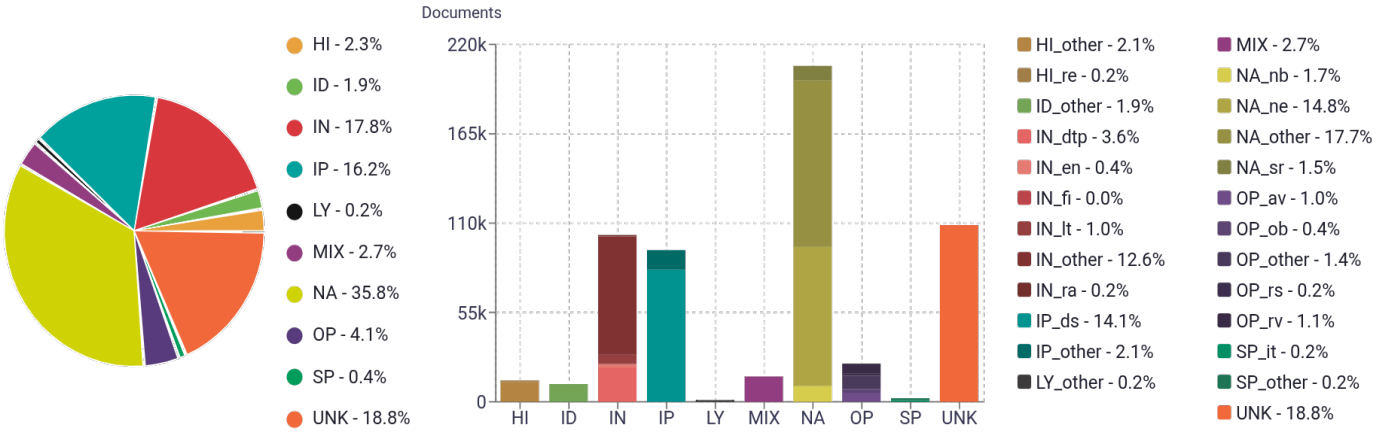
Top 10 domains

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%

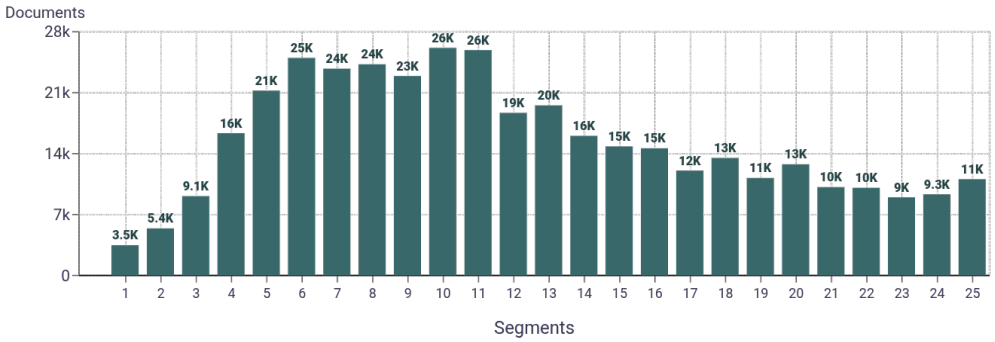
Register labels



MT:5.9% | 34K Documents

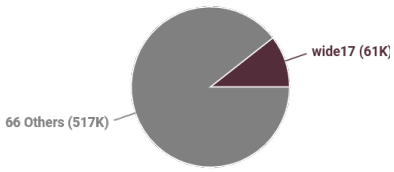
Documents size (in segments)

≤ 25 segments **66.81%** (386K documents)
> 25 segments **33.19%** (192K documents)



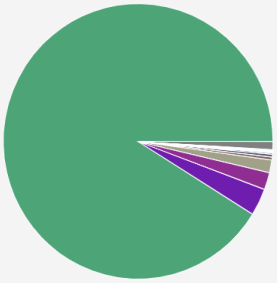
Document collections

CC = 54.95%
IA = 45.05%



Language Distribution

Number of segments in the Chinese (cmn) corpus

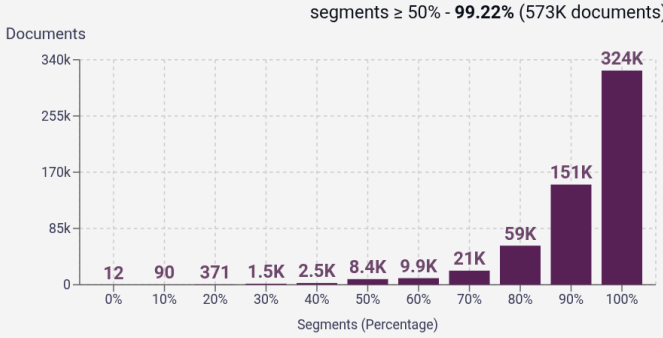


- Chinese (zh) - 15M **(91.1%)**
- Japanese (ja) - 517K **(3.2%)**
- English (en) - 329K **(2.0%)**
- Italian (it) - 242K **(1.5%)**
- French (fr) - 56K **(0.3%)**
- German (de) - 52K **(0.3%)**
- Spanish (es) - 27K **(0.2%)**
- Korean (ko) - 22K **(0.1%)**
- Russian (ru) - 22K **(0.1%)**
- Wu Chinese (wuu) - 19K **(0.1%)**
- 159 Others - 154K **(0.9%)**

*Chinese (cmn) identification might be inaccurate because it is not supported by FastSpell.

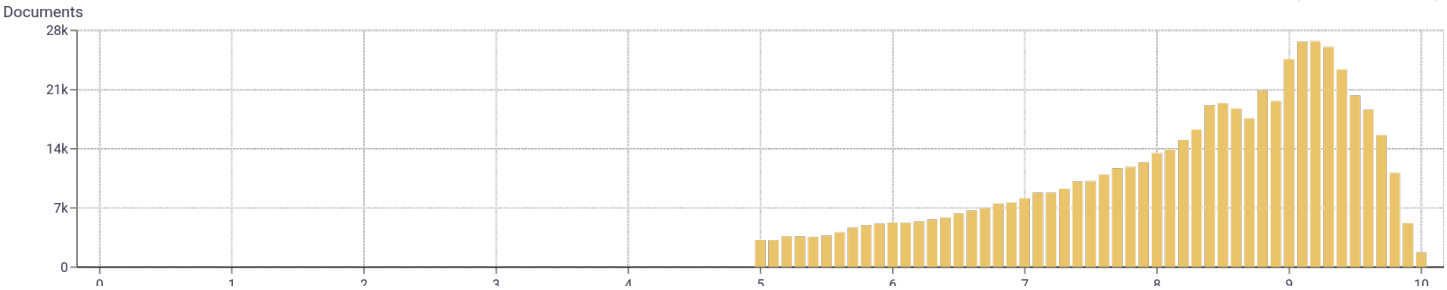
Percentage of segments in Chinese (cmn) inside documents

segments < 50% - **0.78%** (4.5K documents)
segments ≥ 50% - **99.22%** (573K documents)



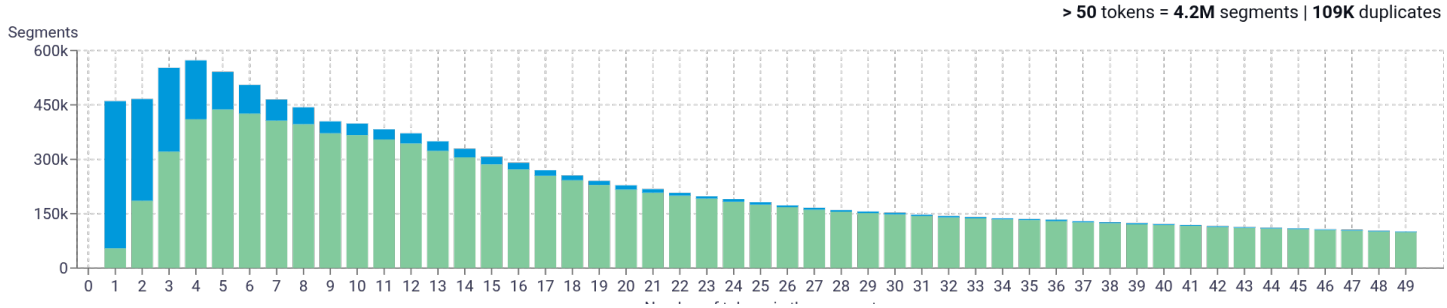
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (578K documents)

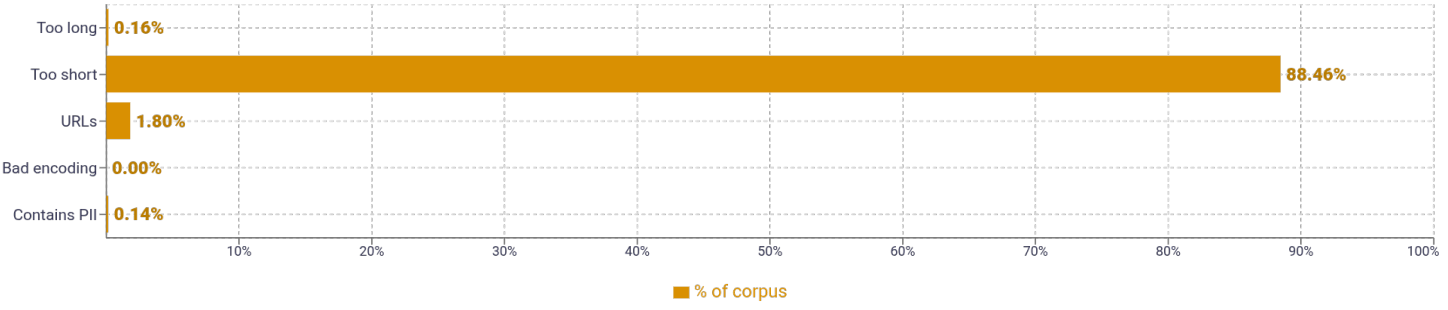


Segment length distribution by token

≤ **49** tokens = **12M** segments | **1.7M** duplicates
> **50** tokens = **4.2M** segments | **109K** duplicates



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 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

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				