

General overview

| Corpus | Date | Language |
|------------------|-----------|------------------|
| hplt-v3-cmn_Hant | 9/18/2025 | Mandarin Chinese |

Volumes

| Docs | Segments | Unique segments | Tokens | Characters | Size |
|-------------|---------------|-------------------------|--------|-----------------|----------|
| 113,442,082 | 2,369,145,108 | 1,480,131,282 (62.48 %) | 115B | 193,066,680,582 | 480.4 GB |

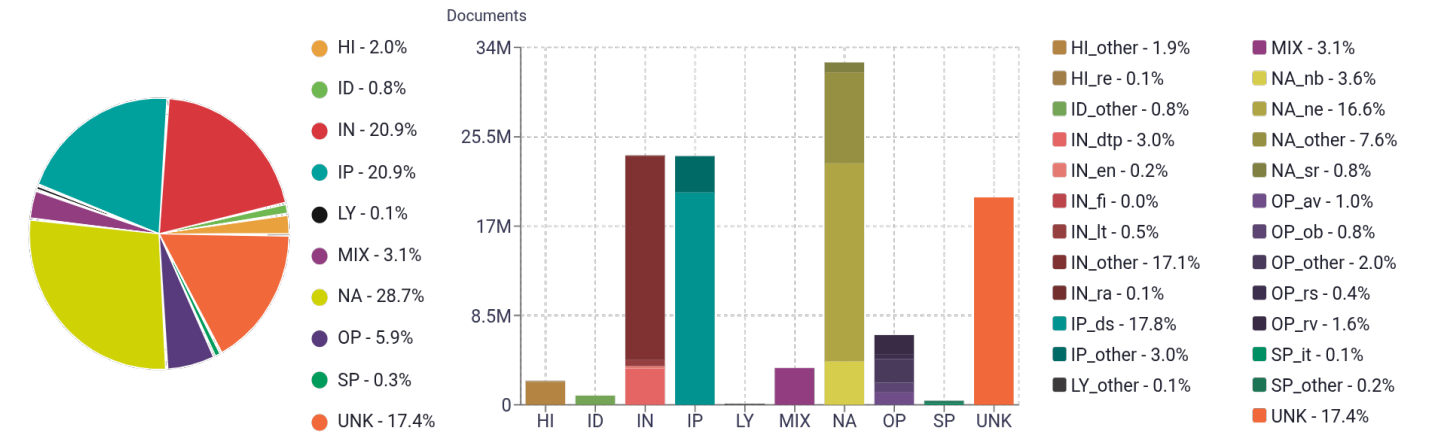
Top 10 domains

| Domain | Docs | % of total |
|-----------------------------------|------|------------|
| pixnet.net | 5.1M | 4.47% |
| songtiankeji1.com | 4.9M | 4.29% |
| b111.net | 1.1M | 0.98% |
| yahoo.com | 1.1M | 0.95% |
| udn.com | 919K | 0.81% |
| pscyhd.com | 808K | 0.71% |
| blogspot.com | 574K | 0.51% |
| hhtdz.com | 548K | 0.48% |
| pchome.com.tw | 540K | 0.48% |
| zjggzs.com | 505K | 0.45% |

Top 10 TLDs

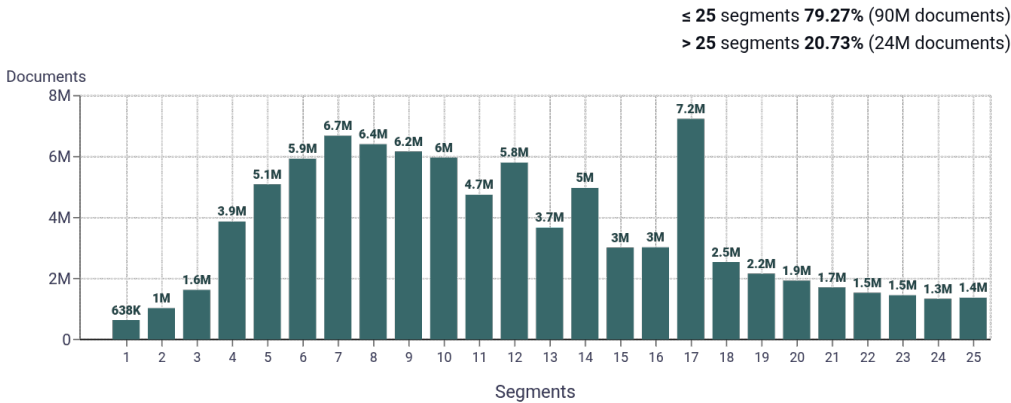
| Domain | Docs | % of total |
|--------|------|------------|
| com | 74M | 65.01% |
| net | 9.6M | 8.50% |
| com.tw | 7.4M | 6.50% |
| tw | 4.5M | 3.94% |
| cn | 2M | 1.79% |
| live | 1.8M | 1.59% |
| org | 1.8M | 1.57% |
| xyz | 1M | 0.90% |
| com.cn | 1M | 0.89% |
| hk | 968K | 0.85% |

Register labels

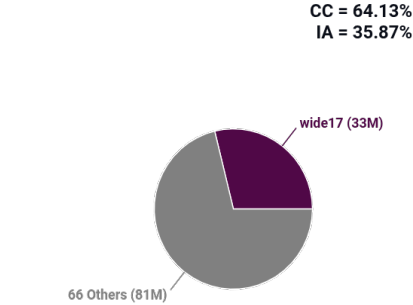


MT:3.9% | 4.4M Documents

Documents size (in segments) ⓘ

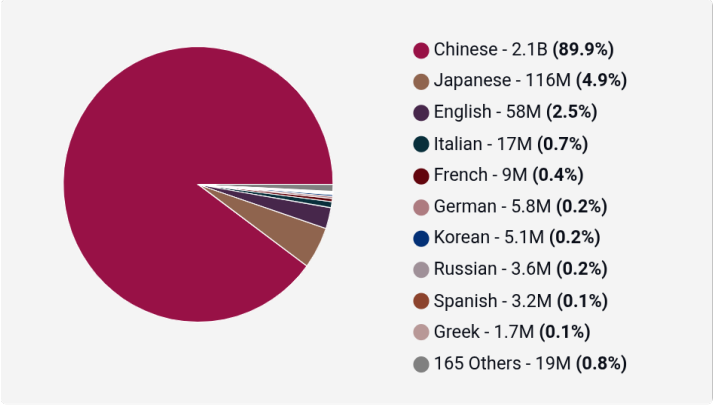


Document collections

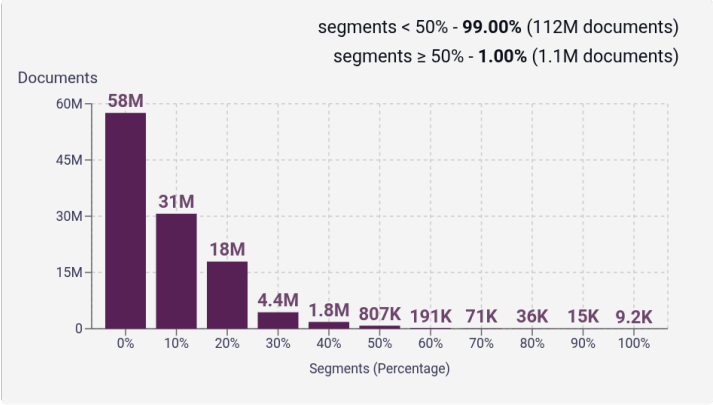


Language Distribution

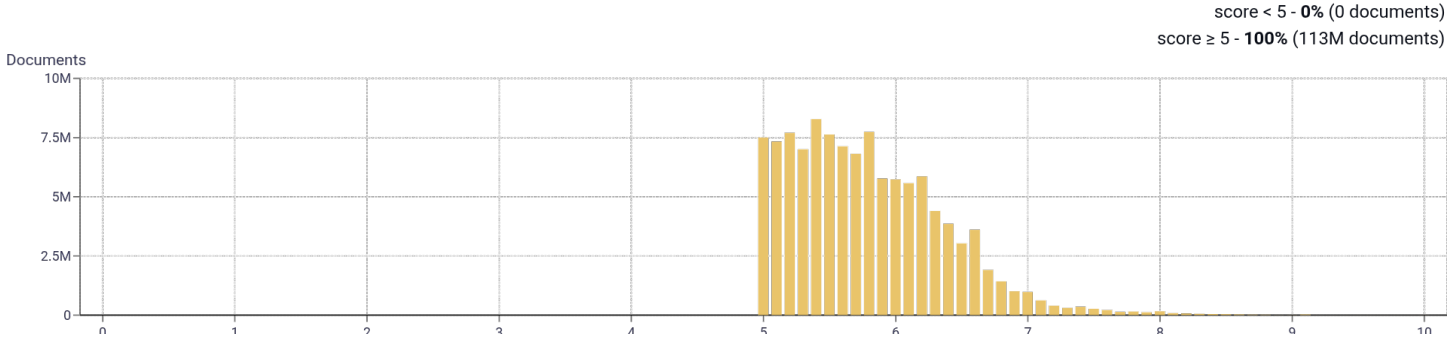
Number of segments in the Mandarin Chinese corpus



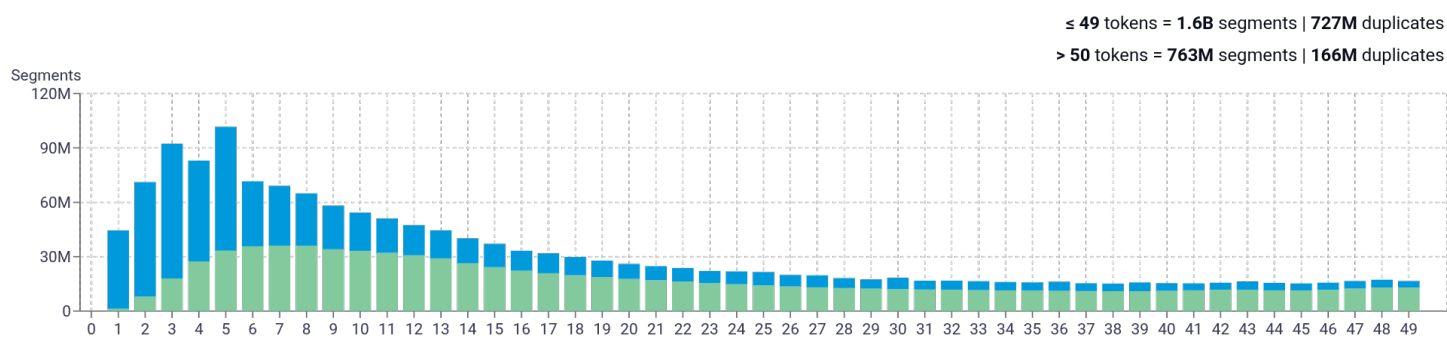
Percentage of segments in Mandarin Chinese inside documents



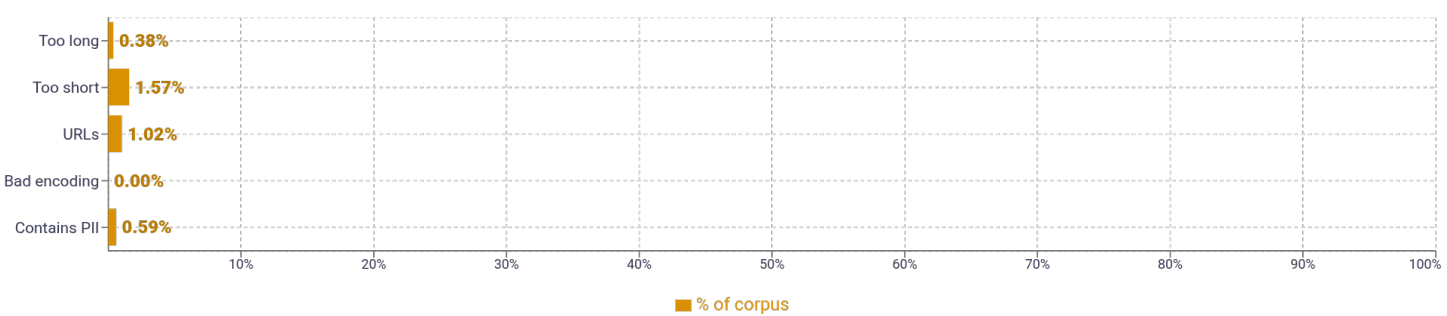
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

| SIZE | N-GRAMS | |
|------|---|--|
| 1 | 不 716,936,943 中 347,076,708 個 334,837,650 都 317,956,846 上 281,898,881 | |
| 2 | 已經 54,634,681 不斷 35,515,815 不會 34,010,080 過程中 31,487,919 最大 27,171,798 | |
| 3 | 越來越 19,764,693 靈活多變 9,829,478 這一內容 9,741,064 產品的系統 9,740,925 便是環保 9,739,224 | |
| 4 | 最突出的優勢 9,739,897 小編也有介紹 9,738,757 線棒之所以廣泛應用 9,738,648 優勢便是環保 9,738,639 環保以及結構輕巧 9,738,633 | |
| 5 | 小編也有介紹過 9,738,679 最突出的優勢便 9,738,648 便是環保以及結構 9,738,633 這一內容前面小編 9,738,632 廣泛應用於各個 6,060,421 | |

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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>, , , 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 | | | | |