# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Analytics date | Language |
|---|---|---|
| HPLT-v2-zsm_Latn.tsv | 9/22/2024 | Malay (zsm) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Size | Characters |
|---|---|---|---|---|---|
| 18,416,407 | 579,818,034 | | | 72.88 GB | 77,866,047,125 |

## Top 10 domains

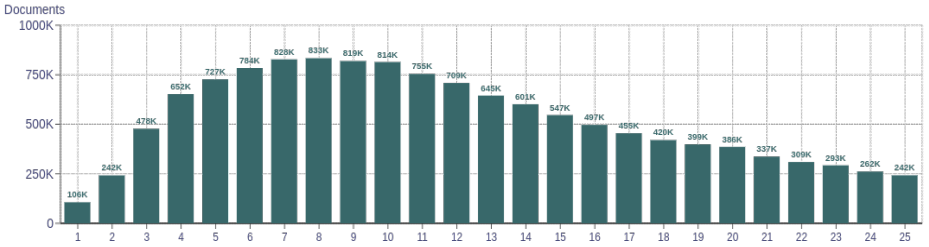
| Domain | Docs | % of total |
|---|---|---|
| blogspot.com | 7.4M | 40.40 |
| blogspot.my | 1.1M | 6.15 |
| wordpress.com | 337K | 1.83 |
| blogspot.sg | 278K | 1.51 |
| wikipedia.org | 254K | 1.38 |
| hotels.com | 128K | 0.69 |
| mstar.com.my | 88K | 0.48 |
| sinarharian.com.my | 88K | 0.48 |
| amanz.my | 82K | 0.44 |
| utusan.com.my | 80K | 0.44 |

## Top 10 TLDs

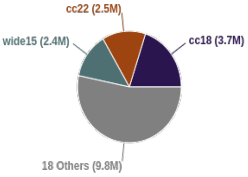
| Domain | Docs | % of total |
|---|---|---|
| com | 13M | 69.96 |
| my | 2M | 10.68 |
| com.my | 714K | 3.88 |
| org | 440K | 2.39 |
| net | 411K | 2.23 |
| sg | 303K | 1.65 |
| gov.my | 161K | 0.87 |
| info | 159K | 0.86 |
| edu.my | 86K | 0.46 |
| com.au | 52K | 0.28 |

## Documents size (in segments)

**<= 25** segments **71.35%** (13M documents)
**> 25** segments **28.65%** (5.3M documents)

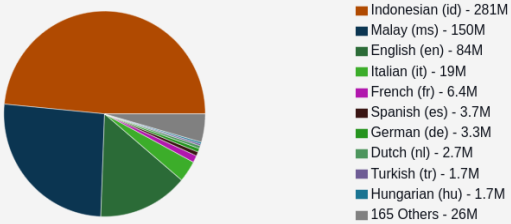



## Documents by collection



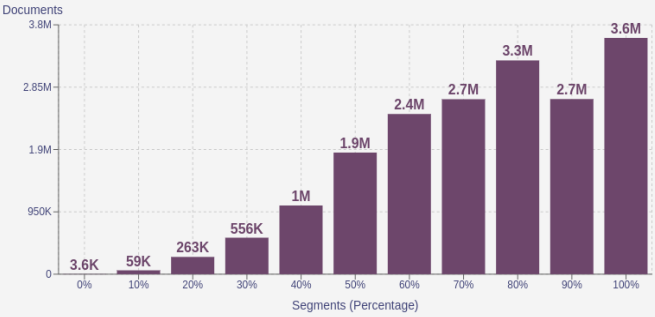

## Language Distribution

### Number of segments



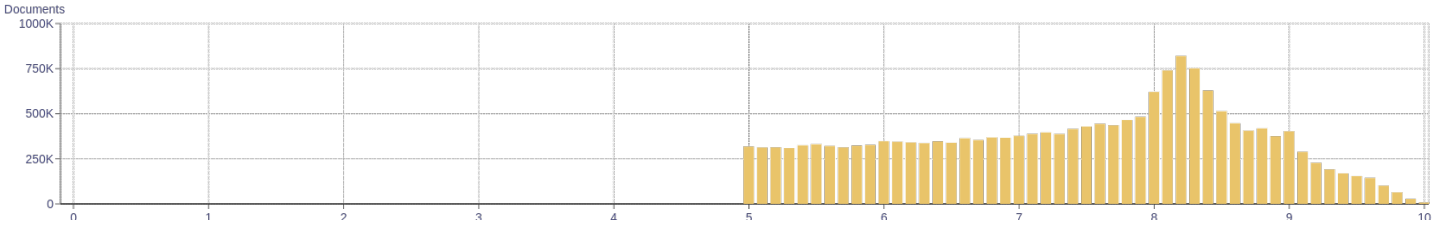

*Malay (zsm) identification might be inaccurate because language is not supported by Fasttext

### Percentage of segments in Malay (zsm) inside documents

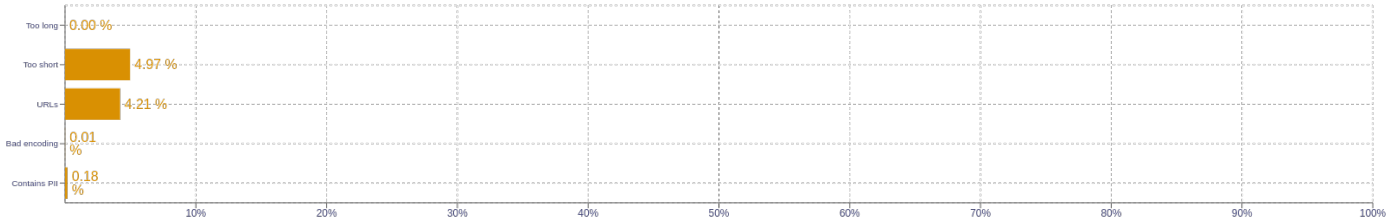



## Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (18M documents)




## 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 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