# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Analytics date | Language |
|---|---|---|
| ml_1.jsonl.tsv | 3/23/2024 | Malayalam (ml) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Size | Characters |
|---|---|---|---|---|---|
| 469,980 | 57,033,693 | 12,708,568 (22.28 %) | 633M | 9.96 GB | |

## Top 10 domains

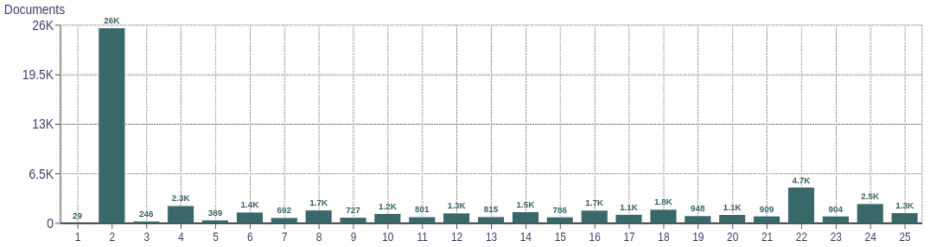
| Domain | Docs | % of total |
|---|---|---|
| blogspot.in | 48K | 10.18 |
| samayam.com | 17K | 3.54 |
| blogspot.com | 13K | 2.86 |
| blogspot.ae | 8.6K | 1.83 |
| wikipedia.org | 8K | 1.70 |
| deepika.com | 6.5K | 1.38 |
| kalapremidaily.com | 5.2K | 1.11 |
| blogspot.co.at | 5.1K | 1.07 |
| fanport.in | 5K | 1.07 |
| asianetnews.com | 4.6K | 0.98 |

## Top 10 TLDs

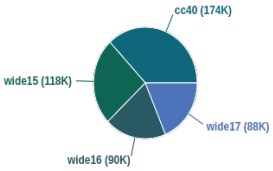
| Domain | Docs | % of total |
|---|---|---|
| com | 301K | 63.97 |
| in | 97K | 20.55 |
| org | 22K | 4.62 |
| ae | 8.7K | 1.86 |
| co.at | 5.1K | 1.07 |
| ie | 4.7K | 1.01 |
| net | 4.5K | 0.95 |
| ch | 3.1K | 0.66 |
| co.uk | 2.5K | 0.53 |
| news | 2.2K | 0.46 |

## Documents size (in segments)

**<= 25** segments **12%** (56K documents)
**> 25** segments **88%** (414K documents)

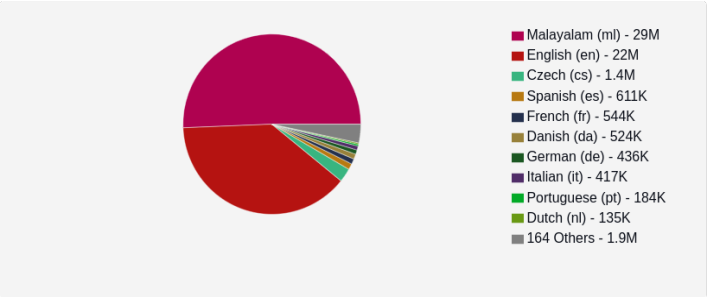



## Documents by collection

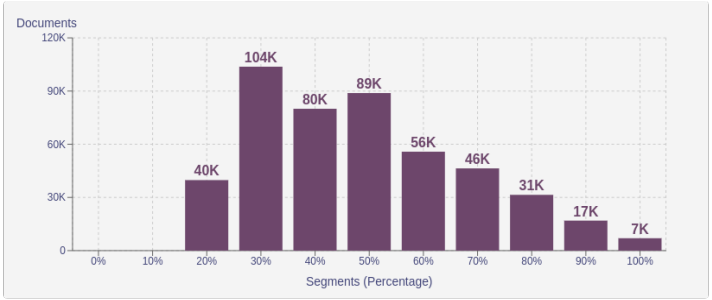



## Language Distribution

### Number of segments

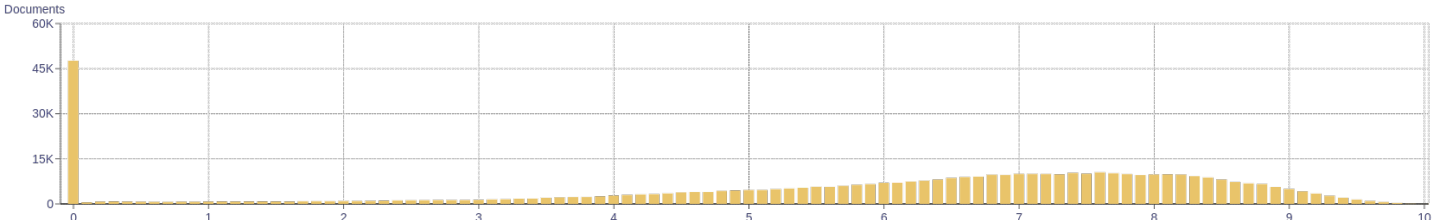



### Percentage of segments in Malayalam (ml) inside documents



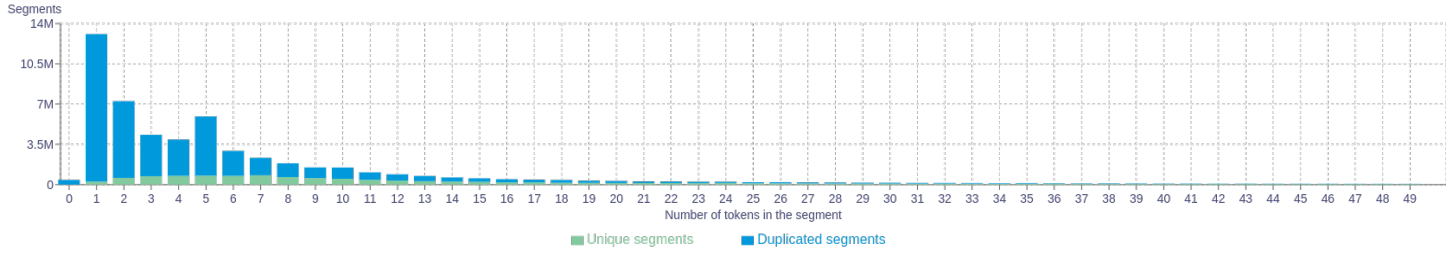

## Distribution of documents by document score

score <= 5 - **27.47%** (129K documents)
score > 5 - **72.53%** (341K documents)

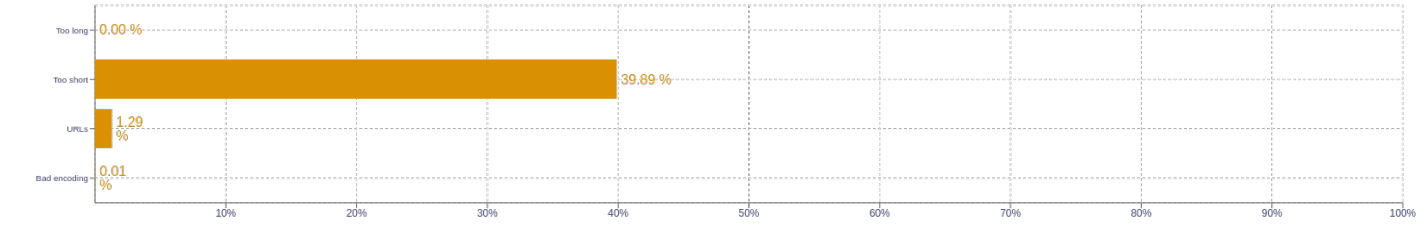



## Segment length distribution by token

**<= 49** tokens = **11M** segments | **44M** duplicates
**> 50** tokens = **2M** segments | **622K** duplicates




## Segment noise distribution



### Frequent n-grams

| Size | n-grams |
|---|---|
| 1 | ഒരു \| 3754966 · ഈ \| 2090109 · to \| 1762323 · the \| 1650615 · news \| 1601203 |
| 2 | span style= \| 275173 · read more \| 265772 · posted by \| 258577 · of the \| 203874 · about us \| 194597 |
| 3 | to twittershare to \| 177459 · share to twittershare \| 177459 · twittershare to facebookshare \| 172892 · to facebookshare to \| 172892 · facebookshare to pinterest \| 172892 |
| 4 | share to twittershare to \| 177459 · twittershare to facebookshare to \| 172892 · to twittershare to facebookshare \| 172892 · to facebookshare to pinterest \| 172892 · links to this post \| 78281 |
| 5 | twittershare to facebookshare to pinterest \| 172892 · to twittershare to facebookshare to \| 172892 · share to twittershare to facebookshare \| 172892 · ൽ പങ്കിടുകfacebook ൽ പങ്കിടുകപിന്റെറസ്റ്റിൽ പങ്കിടുക \| 36035 · twitter ൽ പങ്കിടുകfacebook ൽ പങ്കിടുകപിന്റെറസ്റ്റിൽ \| 36035 |

## 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