HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
mal_Mlym.jsonl.tsv	9/21/2024	Malayalam (ml)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
3,104,759	48,003,517	24,484,289 (51.01 %)	1.2B	23.7 GB	9,443,613,087

Top 10 domains

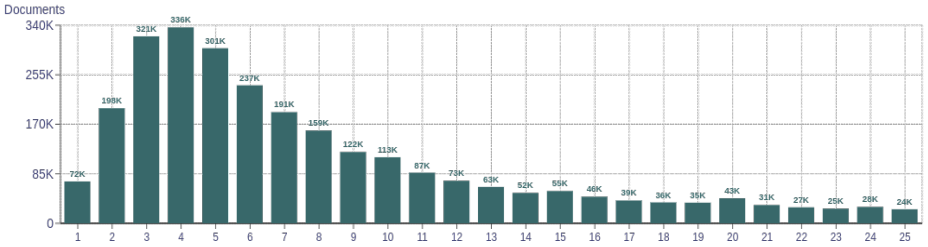
Domain	Docs	% of total
blogspot.com	145K	4.68
wikipedia.org	130K	4.18
thejasnews.com	118K	3.79
mathrubhumi.com	88K	2.82
sirajlive.com	69K	2.23
blogspot.in	65K	2.08
news18.com	55K	1.78
boolokam.com	42K	1.34
manoramaonline.com	41K	1.32
indianexpress.com	40K	1.28

Top 10 TLDs

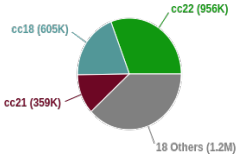
Domain	Docs	% of total
com	2.4M	77.30
in	347K	11.18
org	192K	6.17
net	25K	0.79
ae	15K	0.47
tv	9.9K	0.32
ie	9.5K	0.31
news	9.5K	0.31
gov.in	8.3K	0.27
co.uk	7.2K	0.23

Documents size (in segments)

<= 25 segments **87.49%** (2.7M documents)
> 25 segments **12.51%** (388K documents)

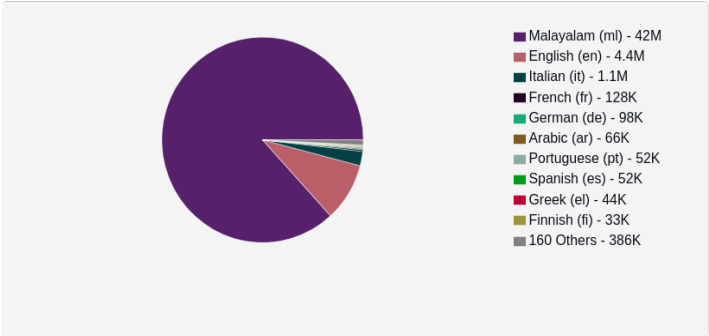


Documents by collection

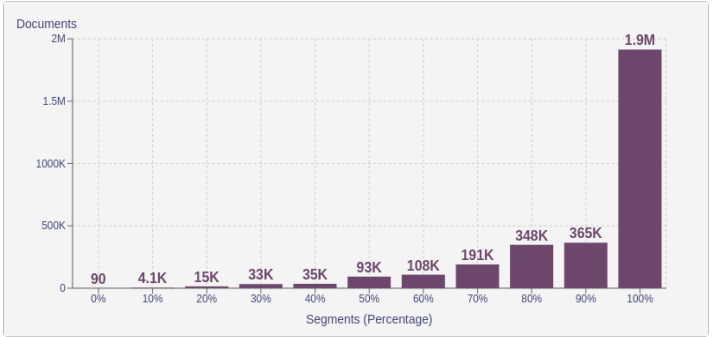


Language Distribution

Number of segments

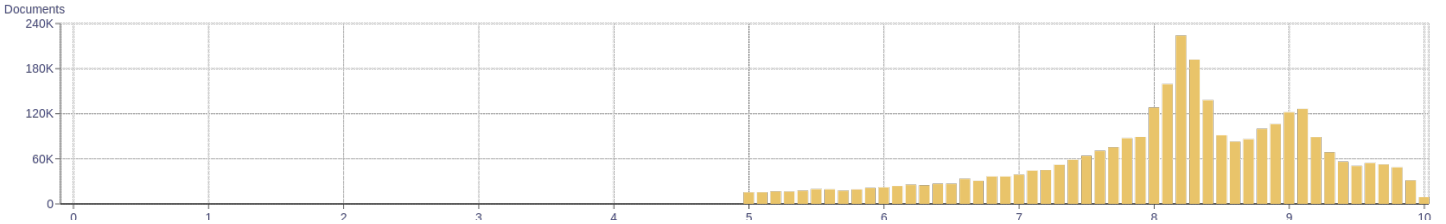


Percentage of segments in Malayalam (ml) inside documents



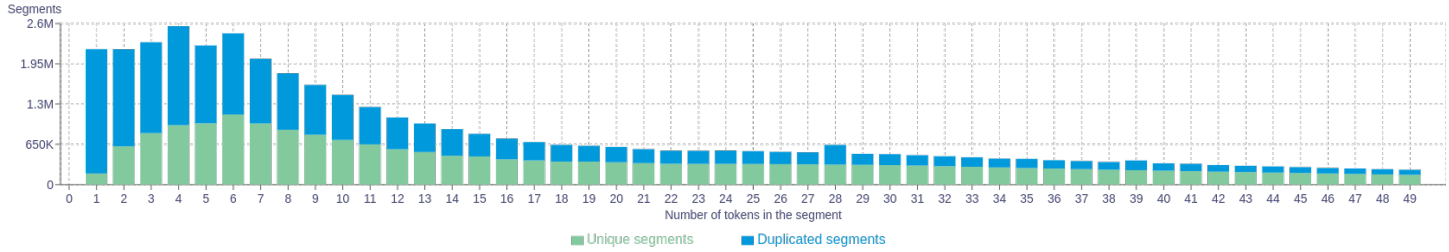
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (3.1M documents)

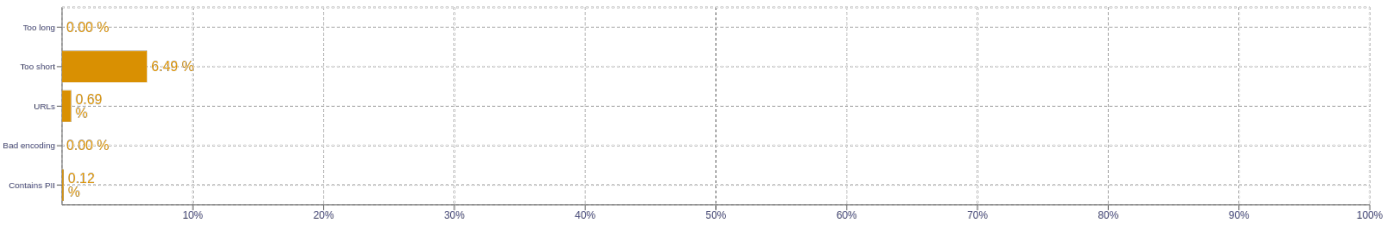


Segment length distribution by token

<= 49 tokens = **20M** segments | **21M** duplicates
> 50 tokens = **6.8M** segments | **2.6M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	പറഞ്ഞു | 2103251 · പി | 1874554 · ഞാന് | 1614324 · കെ | 1614032 · സി | 1539377
2	read more | 244935 · posted by | 177507 · മലയാളത്തിലോ ഇംഗ്ലീഷിലോ | 167197 · വായനക്കാരുടെ അഭിപ്രായങ്ങള് | 141905 · കഴിഞ്ഞ ദിവസം | 138736
3	മലയാളത്തിലോ ഇംഗ്ലീഷിലോ എഴുതുക | 128200 · അവഹേളനപരവും വ്യക്തിപരമായ അധിക്ഷേപങ്ങളും | 119801 · അധിക്ഷേപങ്ങളും അശ്ലീല പദപ്രയോഗങ്ങളും | 111305 · വ്യക്തിപരമായ അധിക്ഷേപങ്ങളും അശ്ലീല | 103018 · വായനക്കാരുടെ അഭിപ്രായങ്ങള് താഴെ | 102810
4	വ്യക്തിപരമായ അധിക്ഷേപങ്ങളും അശ്ലീല പദപ്രയോഗങ്ങളും | 103018 · അവഹേളനപരവും വ്യക്തിപരമായ അധിക്ഷേപങ്ങളും അശ്ലീല | 103018 · വായനക്കാരുടെ അഭിപ്രായങ്ങള് താഴെ എഴുതാവുന്നതാണ് | 102810 · ദയവായി അവഹേളനപരവും വ്യക്തിപരമായ അധിക്ഷേപങ്ങളും | 102604 · വായനക്കാരുടെ അഭിപ്രായ പ്രകടനങ്ങള്ക്കോ അധിക്ഷേപങ്ങള്ക്കോ | 102560
5	അവഹേളനപരവും വ്യക്തിപരമായ അധിക്ഷേപങ്ങളും അശ്ലീല പദപ്രയോഗങ്ങളും | 103018 · വ്യക്തിപരമായ അധിക്ഷേപങ്ങളും അശ്ലീല പദപ്രയോഗങ്ങളും ഒഴിവാക്കുക | 102364 · വായനക്കാരുടെ അഭിപ്രായ പ്രകടനങ്ങള്ക്കോ അധിക്ഷേപങ്ങള്ക്കോ അശ്ലീല | 102364 · ദയവായി അവഹേളനപരവും വ്യക്തിപരമായ അധിക്ഷേപങ്ങളും അശ്ലീല | 102364 · അഭിപ്രായ പ്രകടനങ്ങള്ക്കോ അധിക്ഷേപങ്ങള്ക്കോ അശ്ലീല പദപ്രയോഗങ്ങള്ക്കോ | 102364

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