HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-docslite.ms.tsv	6/9/2024	Malay (ms)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
4,872,339	1,131,639,016	183,095,145 (16.18 %)	12B	52.89 GB	

Top 10 domains

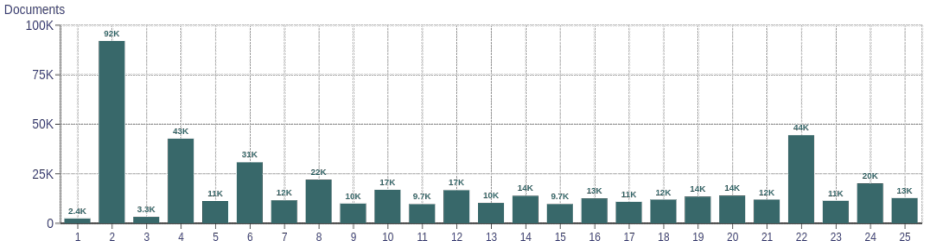
Domain	Docs	% of total
blogspot.my	730K	14.98
blogspot.com	621K	12.75
blogspot.sg	219K	4.49
diebuchsuche.com	91K	1.87
ju8.me	84K	1.73
lyricsparoles.com	68K	1.39
blogspot.in	49K	1.01
blogspot.co.uk	46K	0.95
blogspot.com.au	42K	0.86
wordpress.com	34K	0.70

Top 10 TLDs

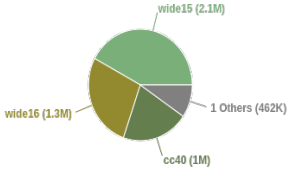
Domain	Docs	% of total
com	2.4M	48.40
my	914K	18.77
sg	220K	4.52
net	137K	2.81
org	136K	2.79
com.my	125K	2.56
me	90K	1.84
in	51K	1.05
co.uk	47K	0.96
com.au	42K	0.86

Documents size (in segments)

<= 25 segments **9.6%** (467K documents)
> 25 segments **90.4%** (4.4M documents)

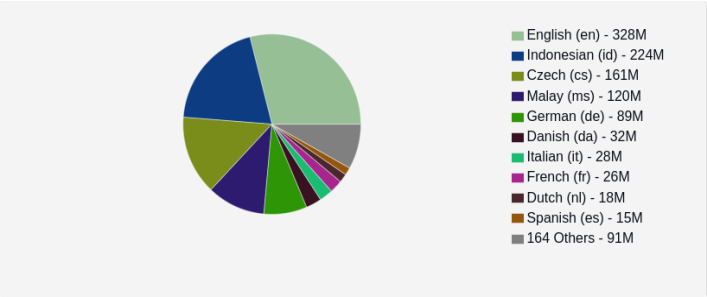


Documents by collection

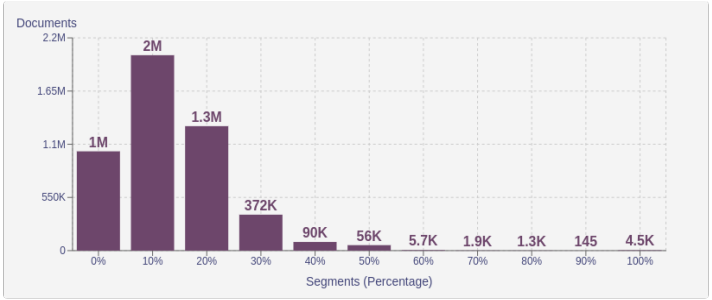


Language Distribution

Number of segments

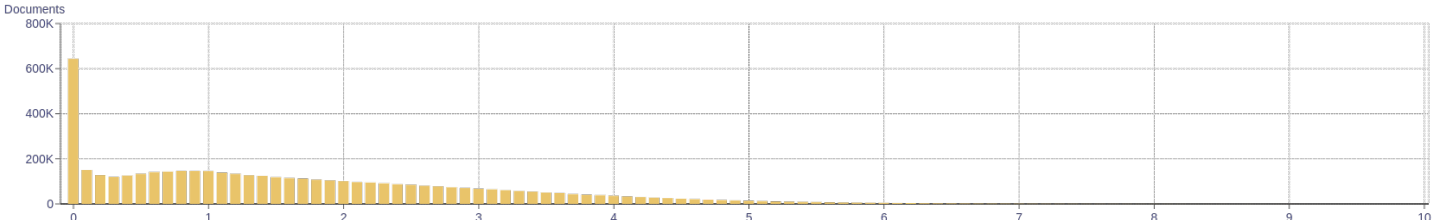


Percentage of segments in Malay (ms) inside documents



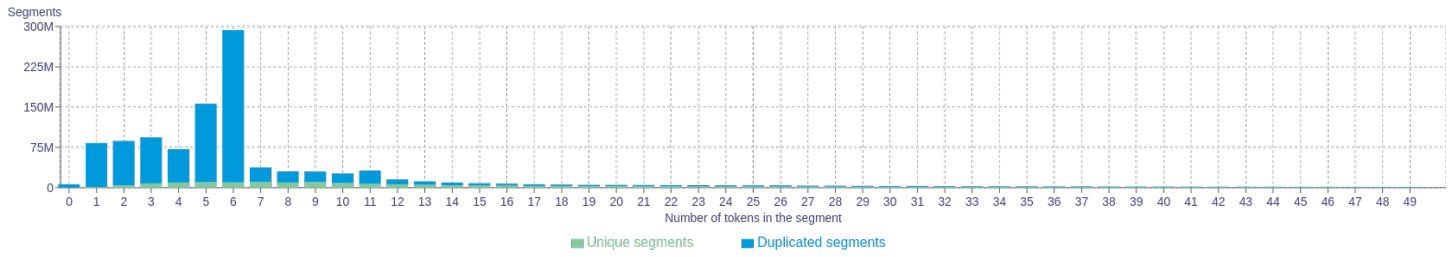
Distribution of documents by document score

score <= 5 - **97.32%** (4.7M documents)
score > 5 - **2.68%** (130K documents)



Segment length distribution by token

<= 49 tokens = **163M** segments | **934M** duplicates
> 50 tokens = **35M** segments | **14M** duplicates



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