HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
lus_Latn.jsonl.tsv	11/5/2024	Mizo (lus)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
160,378	3,433,373	2,006,554 (58.44 %)	146M	624.09 MB	648,740,211

Top 10 domains

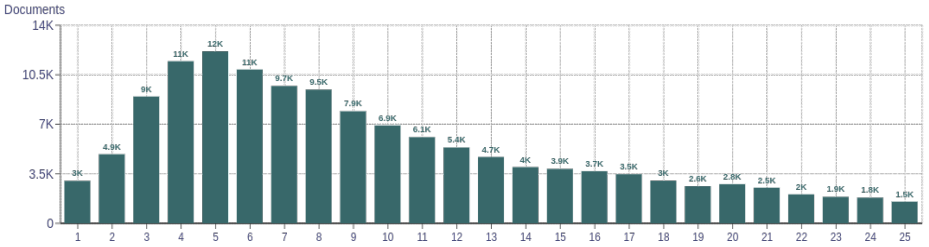
Domain	Docs	% of total
misual.com	19K	11.97
khampat.com	7.8K	4.88
zomidaily.org	6.9K	4.30
blogspot.com	6K	3.74
zothlifim.com	3.8K	2.38
virthli.in	3.7K	2.33
zonet.in	3.6K	2.26
mizolyric.com	3.6K	2.26
timesofmizoram.com	3.4K	2.09
exploremizoram.com	3K	1.86

Top 10 TLDs

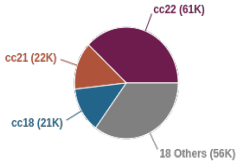
Domain	Docs	% of total
com	97K	60.35
org	21K	13.19
in	15K	9.15
info	7.5K	4.67
net	5.9K	3.68
gov.in	2.1K	1.32
is	1.9K	1.21
no	1.4K	0.86
nic.in	1.3K	0.83
co.in	1.3K	0.81

Documents size (in segments)

<= 25 segments **83.97%** (135K documents)
> 25 segments **16.03%** (26K documents)

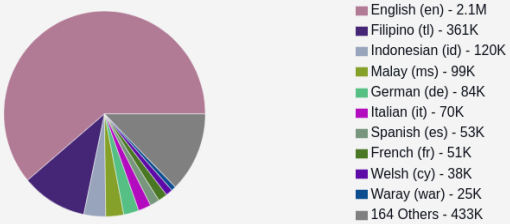


Documents by collection



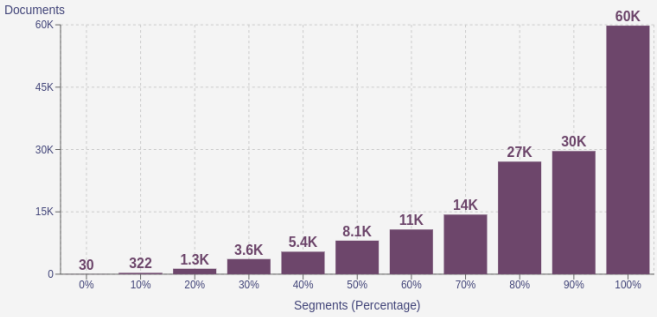
Language Distribution

Number of segments



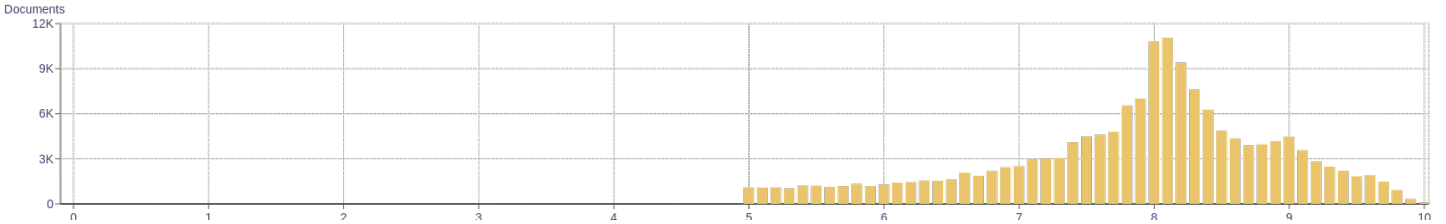
*Mizo (lus) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Mizo (lus) inside documents



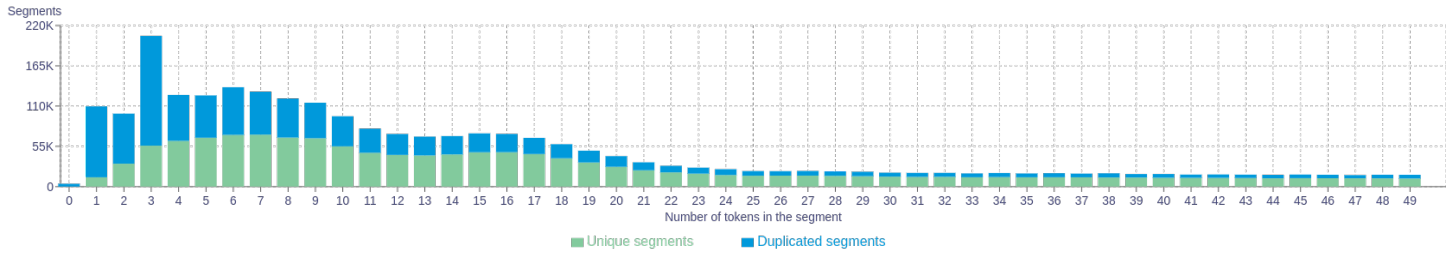
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (160K documents)

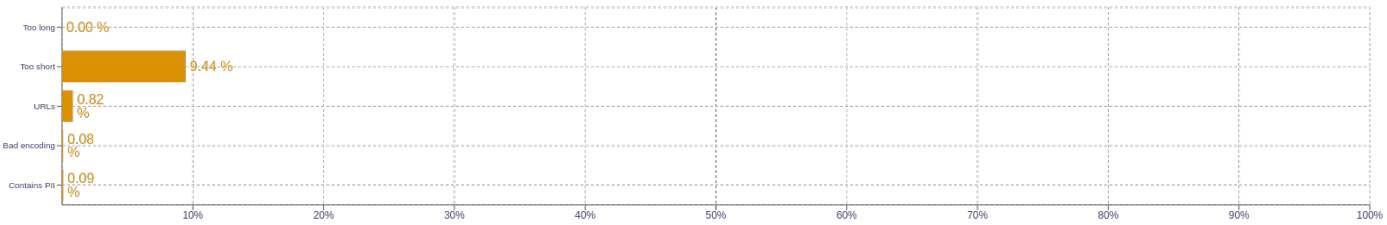


Segment length distribution by token

<= 49 tokens = **1.4M** segments | **1.1M** duplicates
> 50 tokens = **956K** segments | **318K** duplicates



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