HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-docslite.ta.tsv	6/8/2024	Tamil (ta)

Volumes

Docs	Segments	Unique segments	Tokens	Size
1,243,110	217,408,557	124,986 (0.06 %)	2.6B	33.99 GB

Top 10 domains

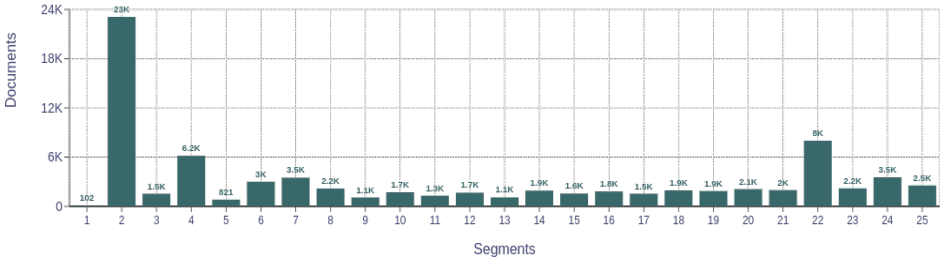
Domain	Docs	% of total
blogspot.in	130K	10.43
blogspot.sg	59K	4.71
blogspot.com	51K	4.10
blogspot.ch	49K	3.96
blogspot.ae	25K	2.02
noolulagam.com	18K	1.48
dinamani.com	16K	1.28
samayam.com	16K	1.28
blogspot.fr	16K	1.27
blogspot.com.au	14K	1.10

Top 10 TLDs

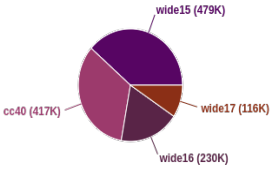
Domain	Docs	% of total
com	671K	54.00
in	192K	15.48
sg	59K	4.71
ch	50K	3.99
org	41K	3.27
net	27K	2.15
ae	25K	2.02
fr	16K	1.28
lk	16K	1.26
ca	15K	1.19

Documents size (in segments)

<= 25 segments **6.3%** (78K documents)
> 25 segments **93.7%** (1.2M documents)

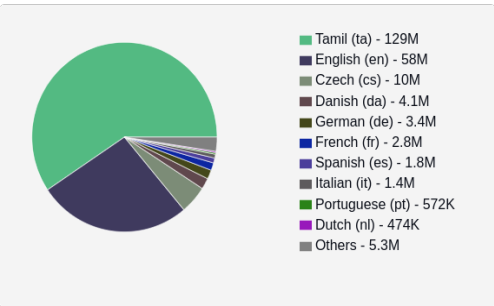


Documents by collection

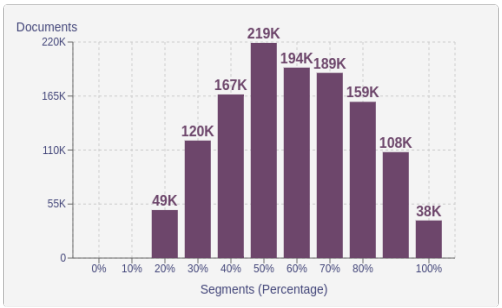


Language Distribution

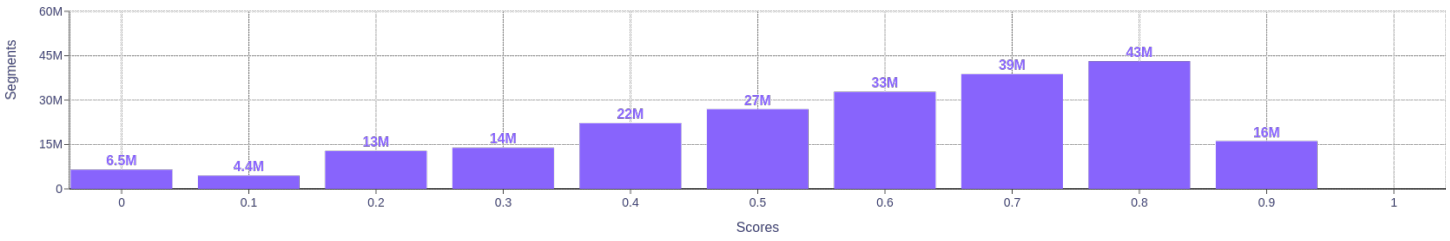
Number of segments



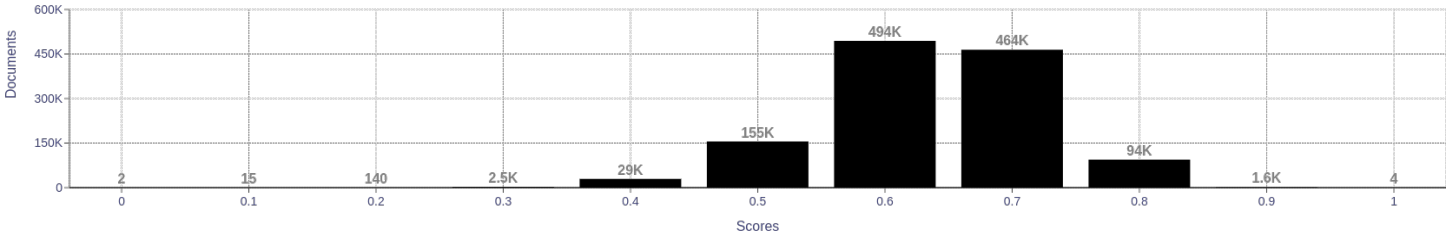
Percentage of segments in Tamil (ta) inside documents



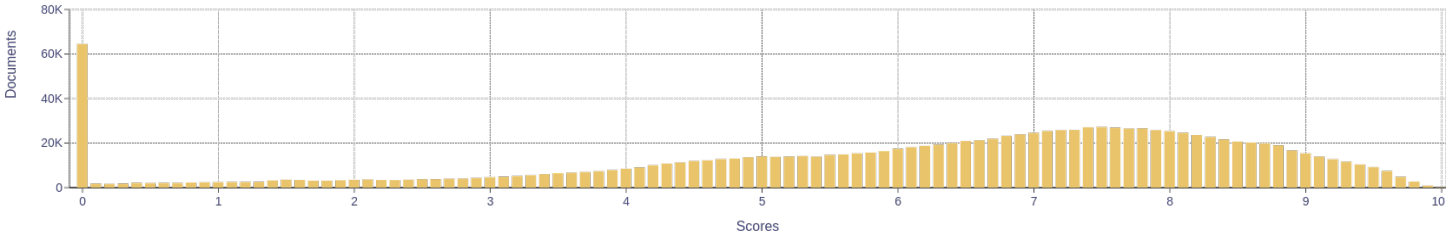
Distribution of segments by fluency score



Distribution of documents by average fluency score



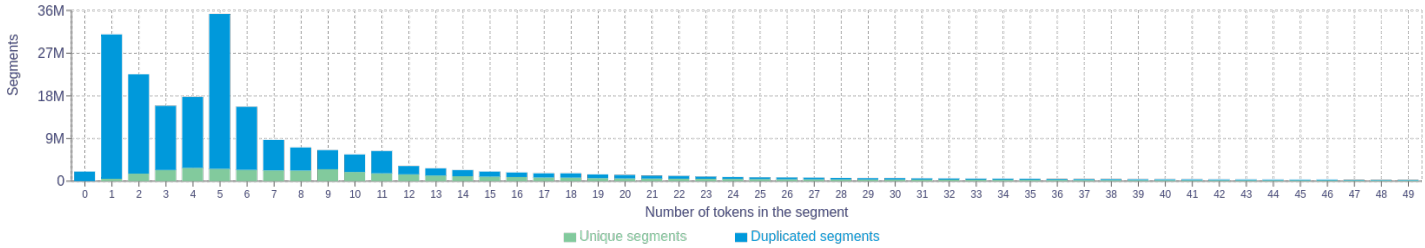
Distribution of documents by document score



Segment length distribution by token

<= 49 tokens = **41M** segments | **169M** duplicates

> 50 tokens = **8M** segments | **3.3M** 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