HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
sin_Sinh.jsonl.tsv	9/17/2024	Sinhala (si)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,152,703	33,707,009	13,612,293 (40.38 %)	934M	11.71 GB	4,948,668,861

Top 10 domains

Domain	Docs	% of total
blogspot.com	220K	19.07
wikipedia.org	28K	2.39
wordpress.com	27K	2.35
baiscopelk.com	26K	2.29
w3lanka.com	17K	1.44
lankacnews.com	15K	1.29
blogspot.com.au	10K	0.91
roar.media	9.5K	0.82
blogspot.kr	8.4K	0.73
hotandfastnews.com	8.3K	0.72

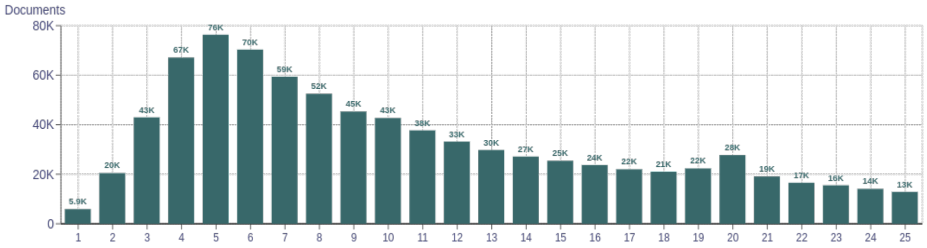
Top 10 TLDs

Domain	Docs	% of total
com	652K	56.52
lk	266K	23.11
org	82K	7.10
net	22K	1.90
info	15K	1.28
com.au	11K	0.96
media	9.5K	0.83
gov.lk	9.4K	0.82
kr	8.4K	0.73
it	7.6K	0.66

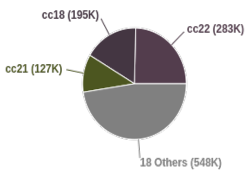
Documents size (in segments)

<= 25 segments **72.12%** (831K documents)

> 25 segments **27.88%** (321K documents)

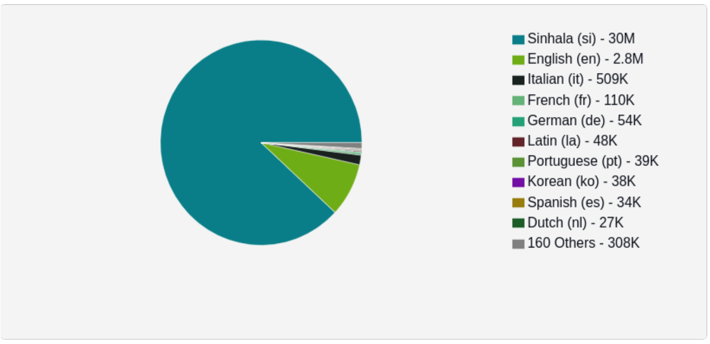


Documents by collection

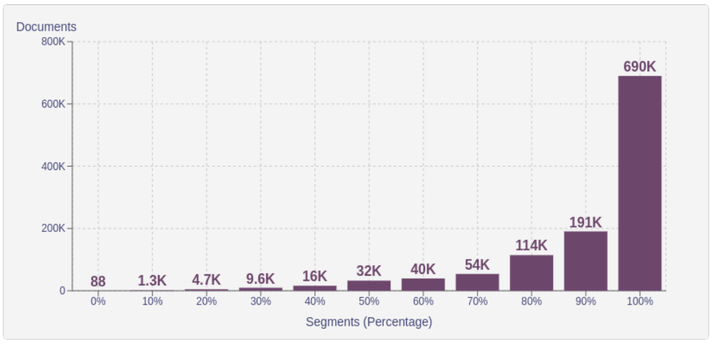


Language Distribution

Number of segments



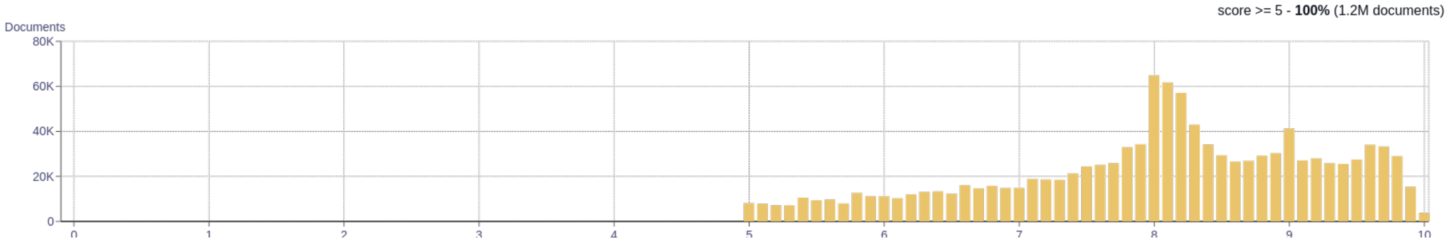
Percentage of segments in Sinhala (si) inside documents



Distribution of documents by document score

score < 5 - **0%** (0 documents)

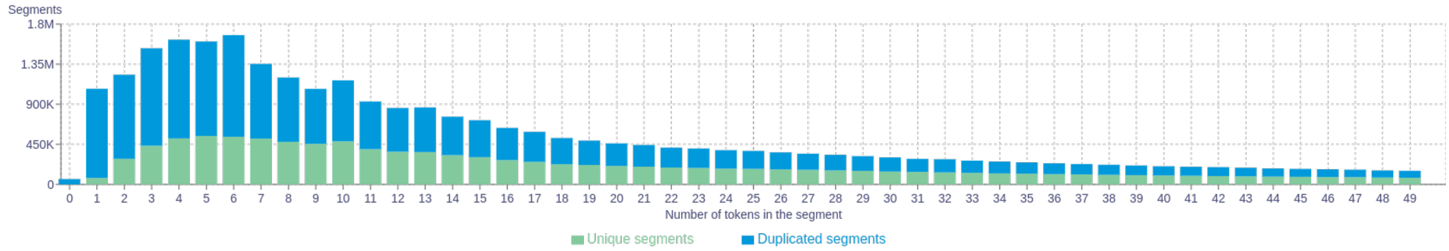
score >= 5 - **100%** (1.2M documents)



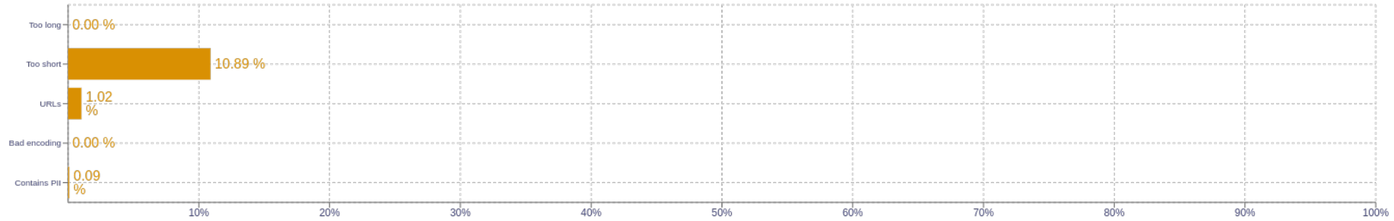
Segment length distribution by token

<= 49 tokens = **11M** segments | **17M** duplicates

> 50 tokens = **5.6M** segments | **2.9M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	එක | 3035192 මම | 2897578 කර | 2473150 කරන | 2375589 මට | 2101839
2	ශ්‍රී ලංකා | 377293 කරන ලද | 263772 කර ඇත | 193077 read more | 188038 කියන එක | 185032
3	ශ්‍රී ලංකා නිදහස් | 44195 ශ්‍රී ලංකා ක්‍රිකට් | 32240 මහින්ද රාජපක්ෂ මහතා | 28640 ජනාධිපති මහින්ද රාජපක්ෂ | 25691 ලබා ගත හැකි | 24312
4	has been removed by | 15888 comment has been removed | 15875 this comment has been | 15694 මූණු පොතට එක් කරන්න | 15510 ඔබේ මූණු පොතට එක් | 15510
5	comment has been removed by | 15870 this comment has been removed | 15693 ඔබේ මූණු පොතට එක් කරන්න | 15510 බූන්දිය ඔබේ මූණු පොතට එක් | 15406 share this boondi on facebook | 15383

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