

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

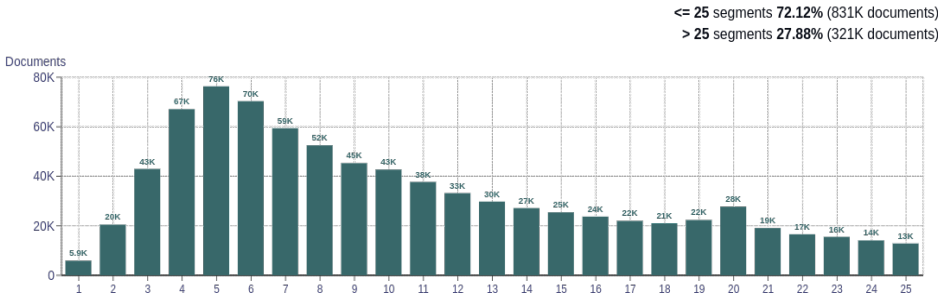
Top 10 domains

Domain	Docs	% of total
blogspot.com	220K	19.07
wikipedia.org	28K	2.39
wordpress.com	27K	2.35
baiscope.lk	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

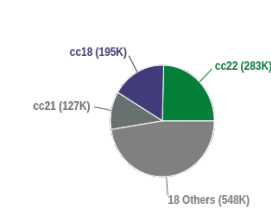
Top 10 TLDs

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)

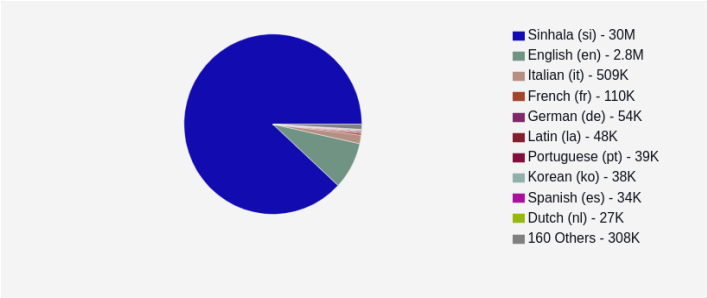


Documents by collection

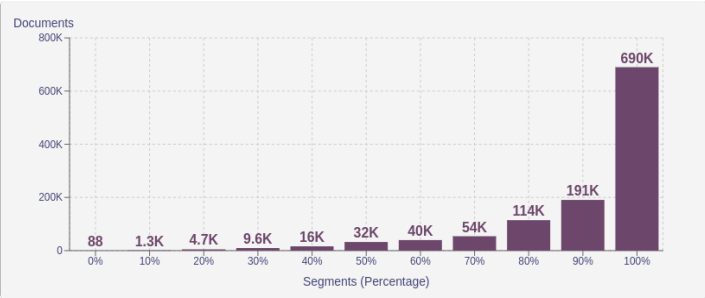


Language Distribution

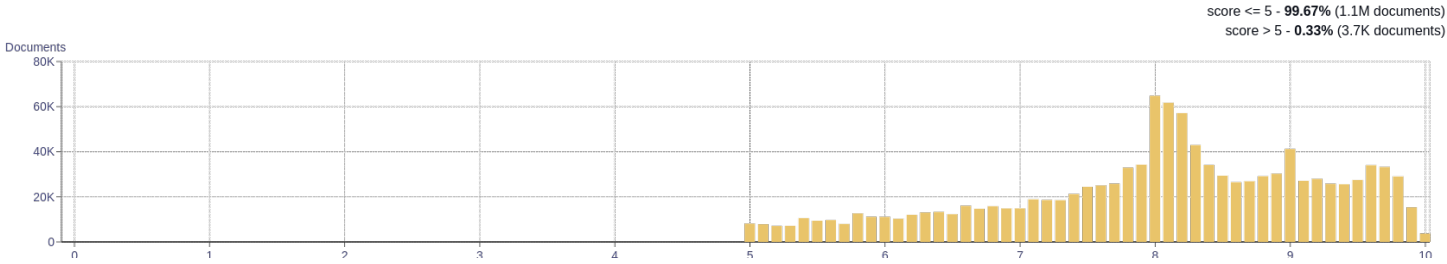
Number of segments



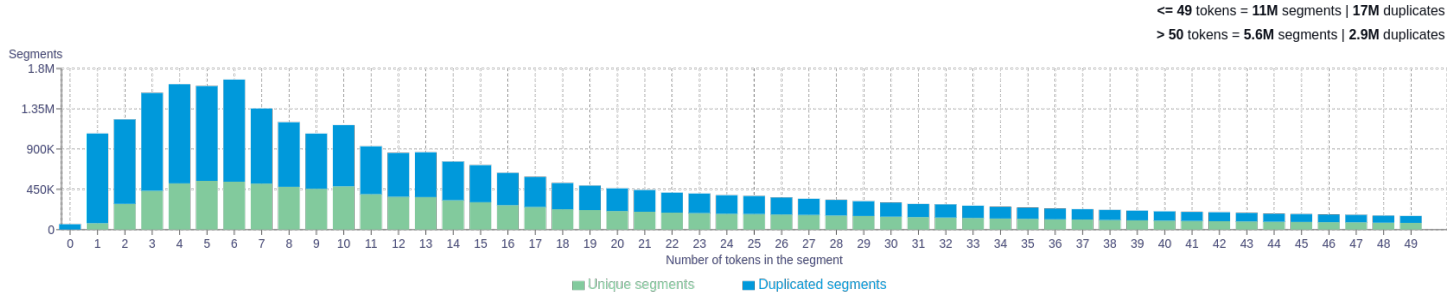
Percentage of segments in Sinhala (si) inside documents



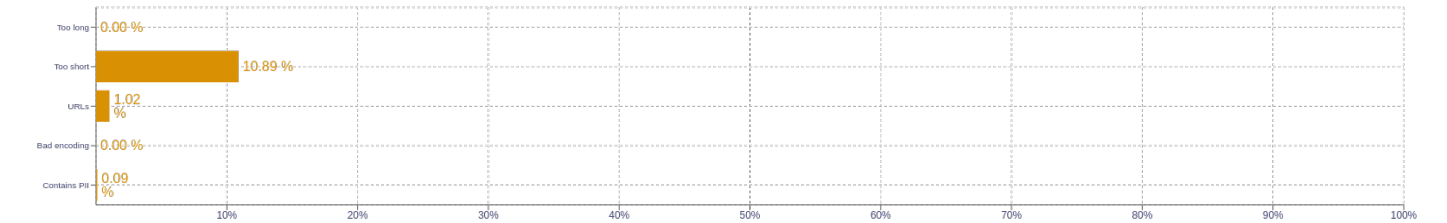
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div><div>එක 3035192</div><div>මම 2897578</div><div>කර 2473150</div><div>කරක 2375589</div><div>මට 2101839</div></div>
2	<div><div>ශ්‍රී ලංකා 377293</div><div>කරක ලද 263772</div><div>කර ඇත 193077</div><div>read more 180038</div><div>කියත එක 185032</div></div>
3	<div><div>ශ්‍රී ලංකා නිදහස් 44195</div><div>ශ්‍රී ලංකා ස්විකට් 32240</div><div>මගින්ද රාජපක්ෂ මහතා 28640</div><div>ජනාධිපති මගින්ද රාජපක්ෂ 25691</div><div>ලබා ගත හැකි 24312</div></div>
4	<div><div>has been removed by 15888</div><div>comment has been removed 15875</div><div>this comment has been 15694</div><div>මුණු පොතට එක් කරන්න 15510</div><div>ඔබේ මුණු පොතට එක් 15510</div></div>
5	<div><div>comment has been removed by 15870</div><div>this comment has been removed 15693</div><div>ඔබේ මුණු පොතට එක් කරන්න 15510</div><div>මුන්ද්‍රිය ඔබේ මුණු පොතට එක් 15406</div><div>share this boondi on facebook 15383</div></div>

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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>