HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
si_1.jsonl.tsv	3/23/2024	Sinhala (si)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
322,515	57,918,718	11,050,808 (19.08 %)	764M	7.45 GB	

Top 10 domains

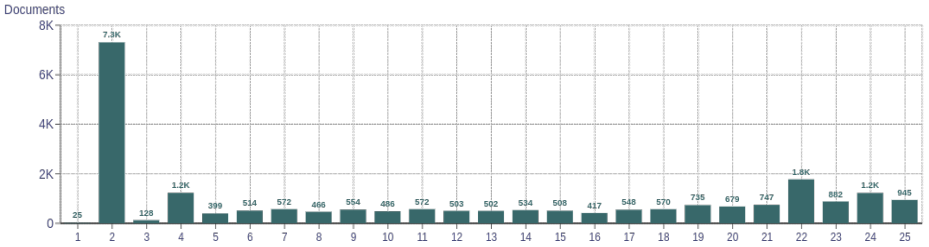
Domain	Docs	% of total
blogspot.com	41K	12.84
baiscopelk.com	7.6K	2.36
blogspot.com.au	6.7K	2.09
w3lanka.com	6.6K	2.05
blogspot.kr	6K	1.85
wordpress.com	5.5K	1.69
blogspot.it	5.3K	1.64
sasrutha.com	5.2K	1.60
lankacnews.com	4.6K	1.43
blogspot.co.il	3.6K	1.12

Top 10 TLDs

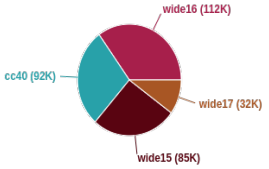
Domain	Docs	% of total
com	167K	51.90
lk	66K	20.36
org	17K	5.16
com.au	6.9K	2.13
kr	6K	1.85
it	5.3K	1.65
info	4.9K	1.52
net	4K	1.24
co.il	3.6K	1.12
ae	3.4K	1.04

Documents size (in segments)

<= 25 segments **7.09%** (23K documents)
> 25 segments **92.91%** (299K documents)

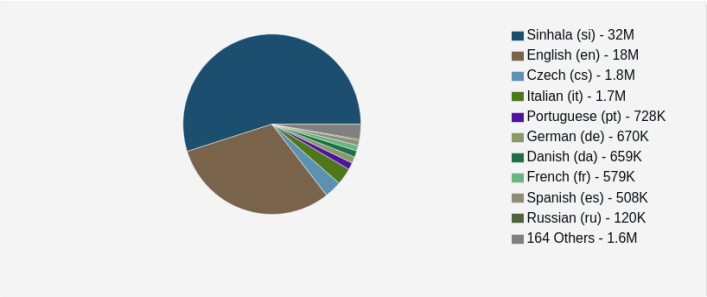


Documents by collection

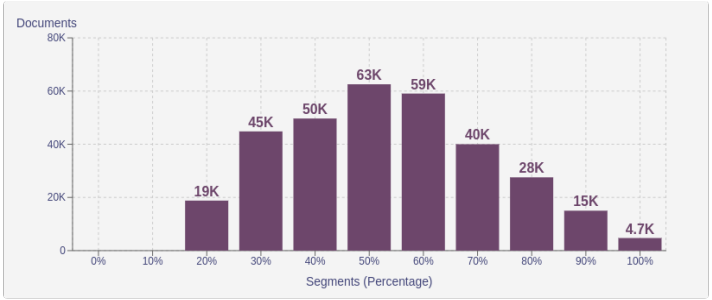


Language Distribution

Number of segments

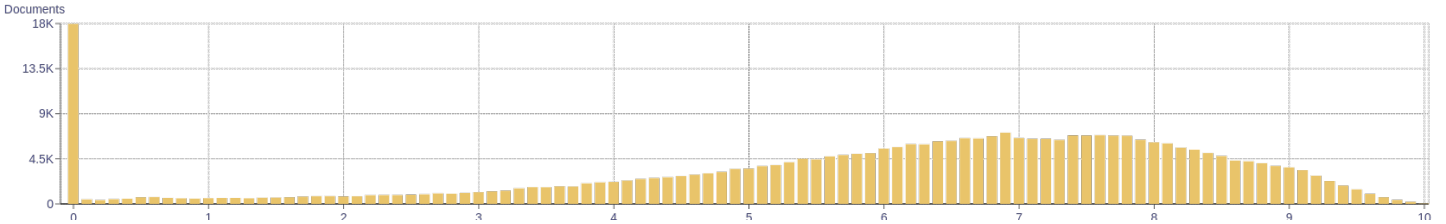


Percentage of segments in Sinhala (si) inside documents



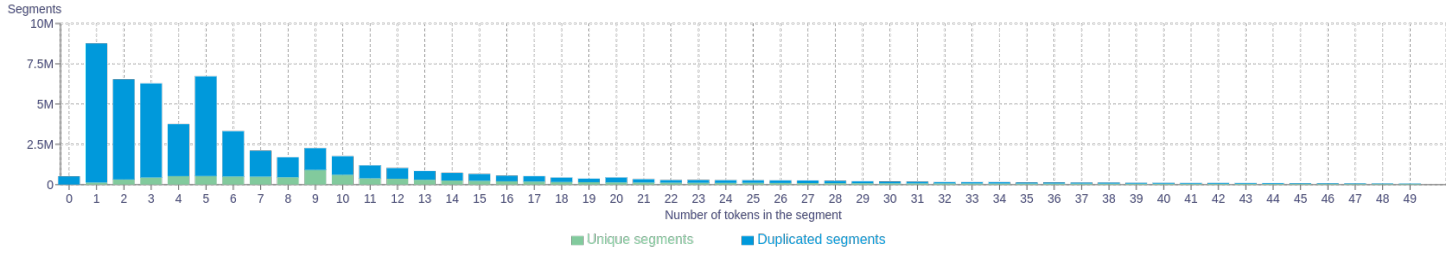
Distribution of documents by document score

score <= 5 - **25.55%** (82K documents)
score > 5 - **74.45%** (240K documents)

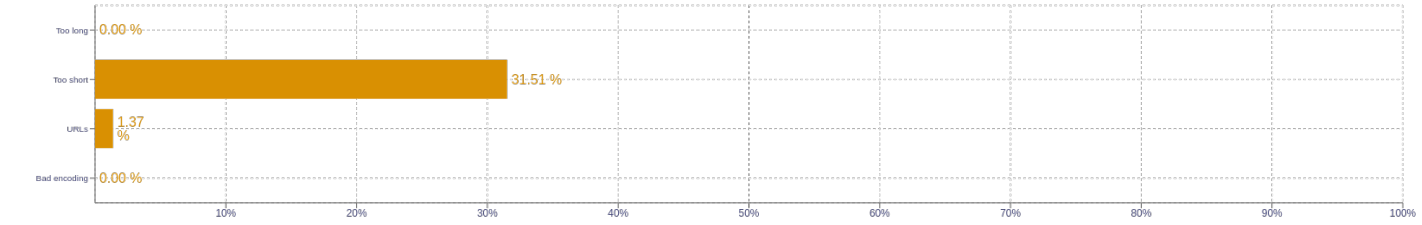


Segment length distribution by token

<= 49 tokens = **9.5M** segments | **46M** duplicates
> 50 tokens = **2.5M** segments | **944K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ago | 2483679 · font | 1962532 · quot | 1806603 · එක | 1736470 · style= | 1732372
2	years ago | 639009 · sinhala subtitles | 555687 · months ago | 546022 · iskoola pota | 424779 · ><span style= | 399335
3	with sinhala subtitles | 362734 · to twittershare to | 216822 · share to twittershare | 216817 · to facebookshare to | 207116 · twittershare to facebookshare | 207099
4	share to twittershare to | 216817 · to twittershare to facebookshare | 207099 · twittershare to facebookshare to | 207096 · to facebookshare to pinterest | 205587 · විෂයට අදාළව ලියන ලදී | 125349
5	to twittershare to facebookshare to | 207096 · share to twittershare to facebookshare | 207095 · twittershare to facebookshare to pinterest | 205587 · තෝමෂයෙ විෂයට අදාළව ලියන ලදී | 104333 · සන්නිවේදන තෝමෂයෙ විෂයට අදාළව ලියන | 101488

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