HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
shn_Mymr.jsonl.tsv	11/4/2024	Shan (shn)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
6,003	92,137	61,045 (66.25 %)	1.6M	55.21 MB	21,124,058

Top 10 domains

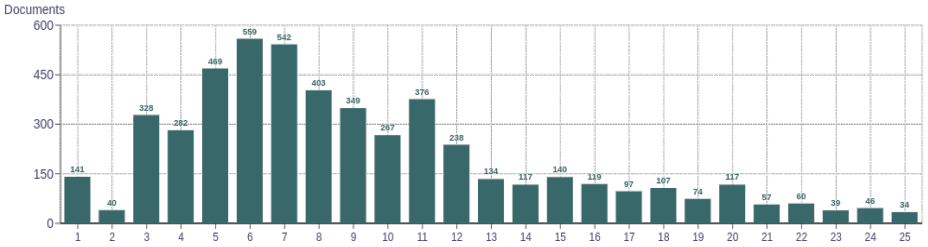
Domain	Docs	% of total
wikipedia.org	1.2K	20.27
taifreedom.com	1K	16.66
panglong.org	950	15.83
shannews.org	788	13.13
bible.is	380	6.33
taisangha.com	188	3.13
taifreedom.net	163	2.72
saisengserk.com	155	2.58
shanhumanrights.org	122	2.03
ssppssa.org	114	1.90

Top 10 TLDs

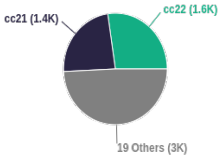
Domain	Docs	% of total
org	3.4K	56.95
com	1.9K	30.92
is	380	6.33
net	182	3.03
media	77	1.28
blog	37	0.62
info	19	0.32
com.mm	9	0.15
com.au	7	0.12
co.uk	5	0.08

Documents size (in segments)

<= 25 segments **85.54%** (5.1K documents)
> 25 segments **14.46%** (867 documents)

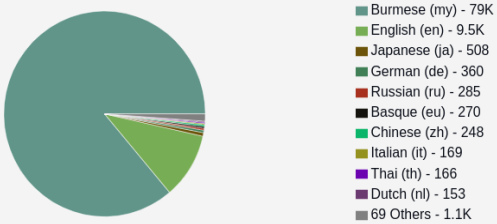


Documents by collection



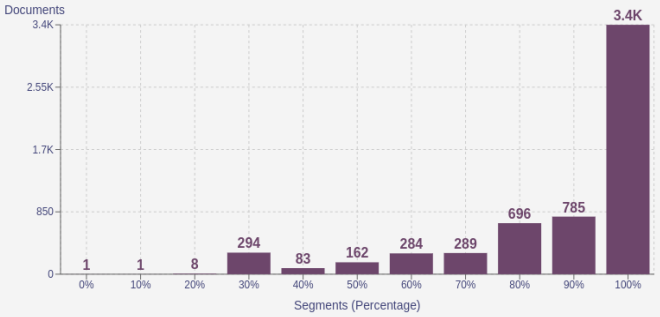
Language Distribution

Number of segments



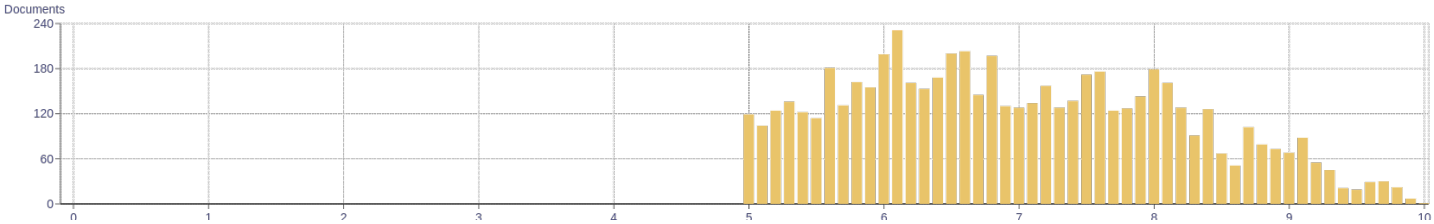
*Shan (shn) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Shan (shn) inside documents



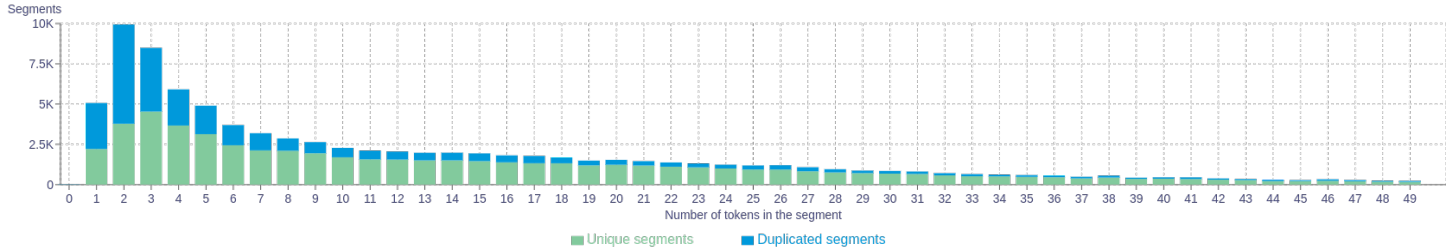
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (6K documents)

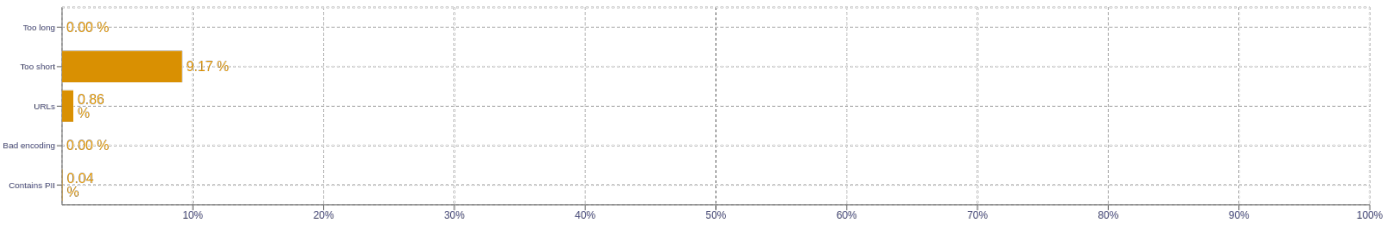


Segment length distribution by token

<= 49 tokens = **58K** segments | **30K** duplicates
> 50 tokens = **4.4K** segments | **1.2K** 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