

General overview

Corpus	Analytics date	Language
shn_Mymr.jsonl.tsv	11/28/2024	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

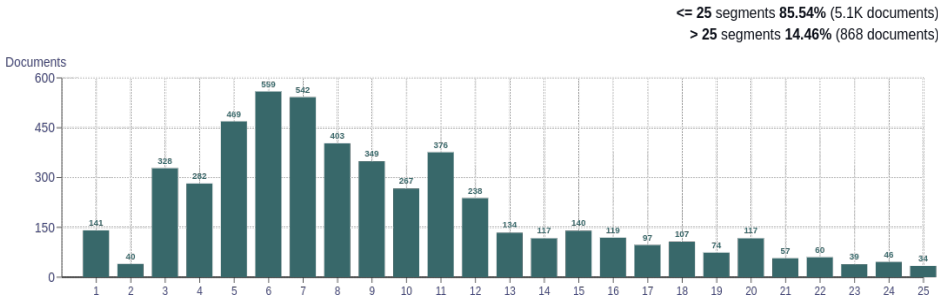
Top 10 domains

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

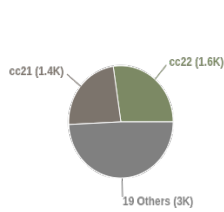
Top 10 TLDs

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)

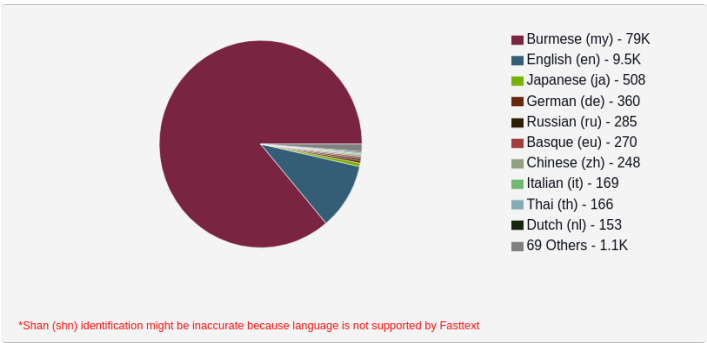


Documents by collection

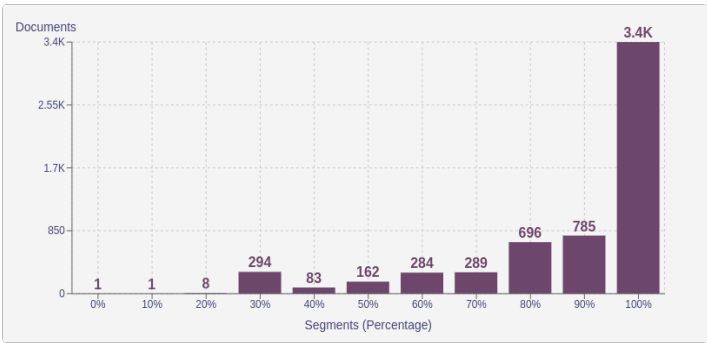


Language Distribution

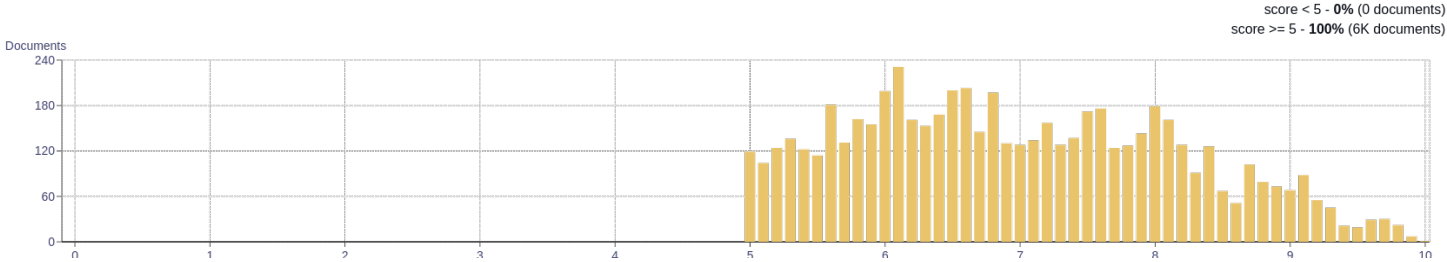
Number of segments



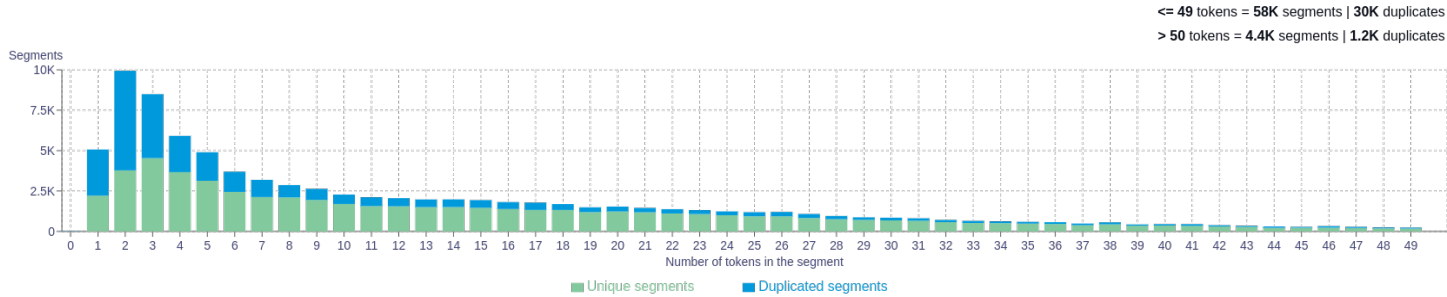
Percentage of segments in Shan (shn) inside documents



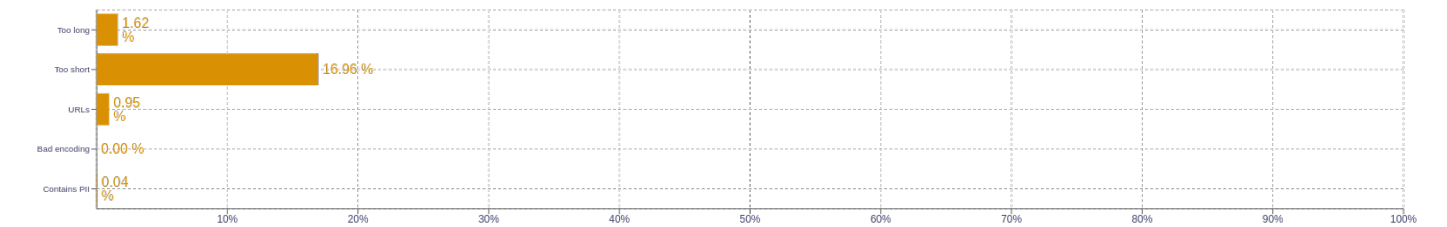
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	သိုင်း   14983 မိုင်း   13657 လှိုင်   8215 ခေ   7888 လွင်   6445
2	တပ်, သိုင်း   1242 လှိုင်, လံး   1170 cov id   1094 writtenbyတံး လွတ်းလံး   1022 တံး လွတ်းလံး   948
3	လှိုင်, လံး မိုင်း   593 တပ်, သိုင်း လိုင်းတံး   472 ှသိကပ်မိုင်း လိုင်းတံး/တပ်, သိုင်း   310 ကခေဂုတ်,ထတ်ဂုမ်းခင်းဂွပ်,မိုင်း ငလးကခေဂွပ်,ဂုမ်း မိုင်း   236 သိုင်း တပ်, သိုင်း   208
4	သိုင်း တပ်, သိုင်း လိုင်းတံး   181 လှိုင် မှ ကြိုကြပ် ခေ   138 ခေ နှင့် ပြည်သူ့ အင်အား   138 မှ ကြိုကြပ် ခေ နှင့်   138 ကြိုကြပ် ခေ နှင့် ပြည်သူ့   138
5	လှိုင် မှ ကြိုကြပ် ခေ နှင့်   138 မှ ကြိုကြပ် ခေ နှင့် ပြည်သူ့   138 ကြိုကြပ် ခေ နှင့် ပြည်သူ့ အင်အား   138 ခေ ၂၀၁၅)သန်း ခေါင်စာရင်း အစီရင်ခံစာ, ၂၀၁၄ ခုနှစ်   136 အစီရင်ခံစာ, ၂၀၁၄ ခုနှစ် လူဦးရေ နှင့် အိမ်ထောင်စု   131

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 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>, <ul>, <ol>, etc.) replaced by newlines.

Language distribution

Language identified with FastSpell (<https://github.com/mbanoni/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>