HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
mya_Mymr.jsonl.tsv	9/18/2024	Burmese (my)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,367,744	30,504,078	14,525,082 (47.62 %)	840M	14.9 GB	5,790,170,855

Top 10 domains

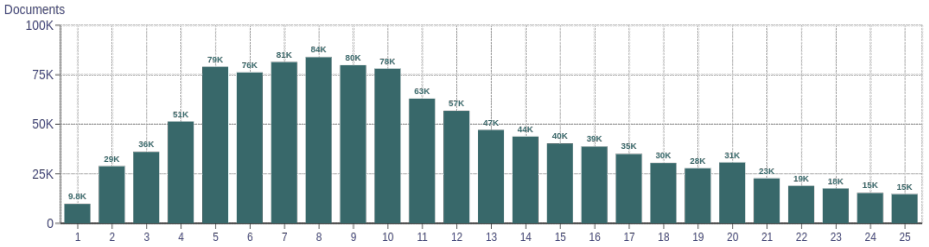
Domain	Docs	% of total
blogspot.com	136K	9.92
dvb.no	67K	4.89
voanews.com	52K	3.83
blogspot.sg	50K	3.64
irrawaddy.com	45K	3.27
wikipedia.org	40K	2.94
thithtoolwin.com	37K	2.70
ygnnews.com	33K	2.41
moemaka.com	33K	2.38
dawnmanhon.com	27K	1.98

Top 10 TLDs

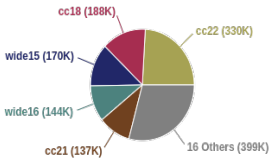
Domain	Docs	% of total
com	831K	60.77
org	122K	8.91
no	91K	6.65
net	70K	5.10
com.mm	52K	3.82
sg	50K	3.64
kr	26K	1.92
gov.mm	24K	1.73
xyz	14K	0.99
ru	6K	0.44

Documents size (in segments)

<= 25 segments **80.96%** (1.1M documents)
> 25 segments **19.04%** (260K documents)

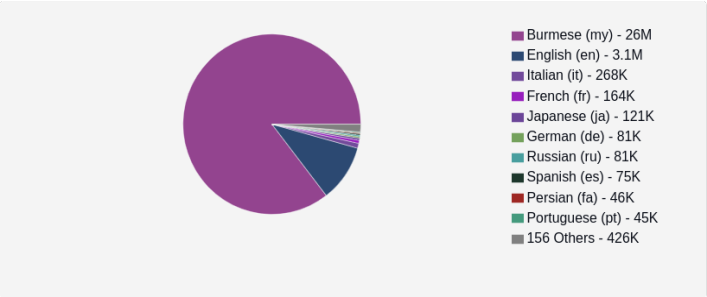


Documents by collection

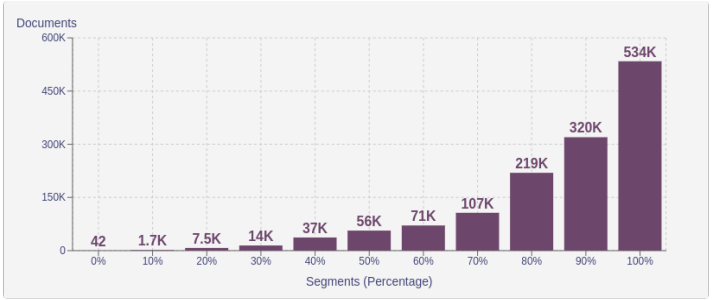


Language Distribution

Number of segments

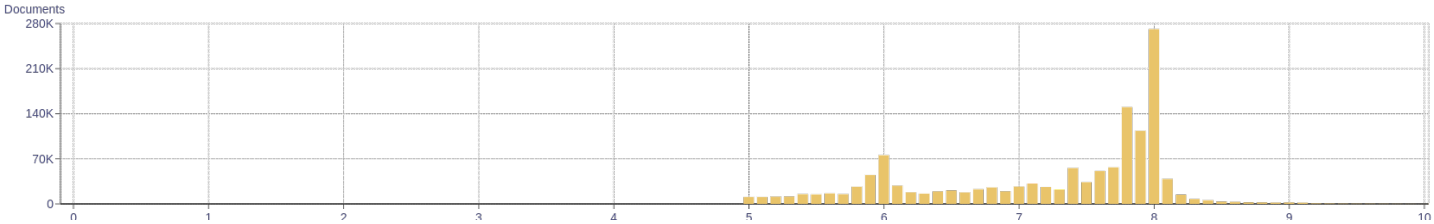


Percentage of segments in Burmese (my) inside documents



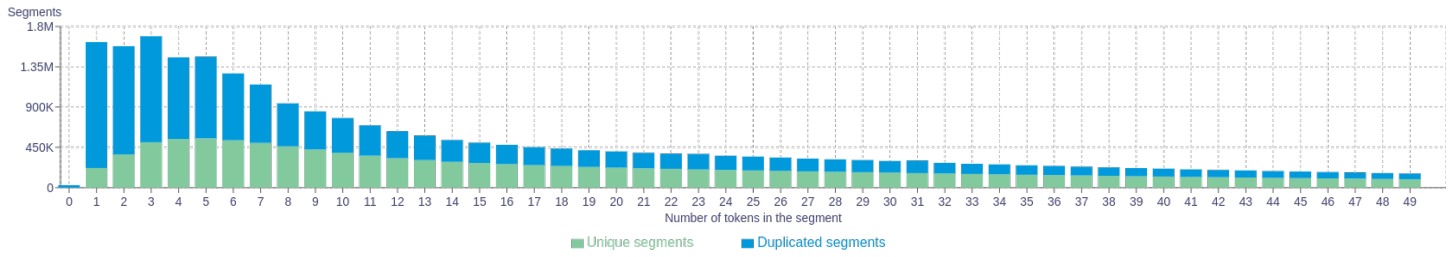
Distribution of documents by document score

score <= 5 - **99.99%** (1.4M documents)
score > 5 - **0.01%** (107 documents)

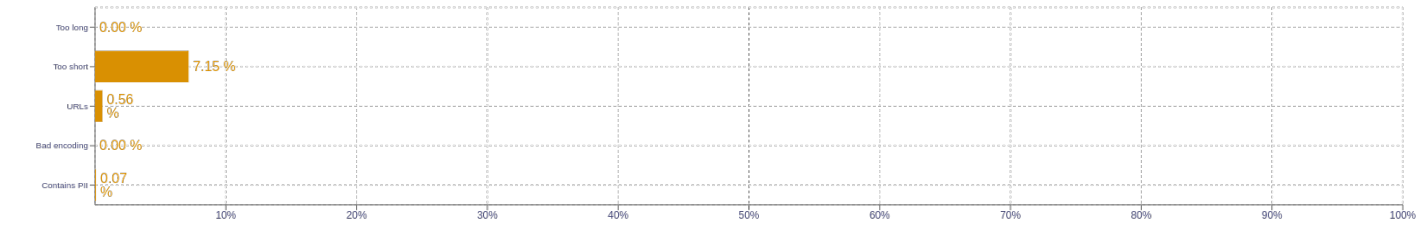


Segment length distribution by token

<= 49 tokens = **12M** segments | **14M** duplicates
> 50 tokens = **5M** segments | **2M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ကို | 19383400 · က | 18664924 · ပါ | 18443866 · ရ | 8481048 · ခဲ့ | 8305887
2	ပါ တယ်။ | 4385521 · ပါ တယ် | 2697883 · ရ ပါ | 1270089 · ခဲ့ ပါ | 1105224 · များ ကို | 998428
3	ဖြစ် ပါ တယ် | 498325 · ခဲ့ ပါ တယ်။ | 441762 · ရ ပါ တယ်။ | 425727 · က ပါ တယ်။ | 370940 · ခဲ့ ပါ တယ် | 289607
4	၎ က ပါ တယ်။ | 115002 · တာ ဖြစ် ပါ တယ် | 95318 · က ကြော ပါ တယ် | 94461 · မှာ ဖြစ် ပါ တယ် | 89085 · ခဲ့ ၎ က ပါ | 82753
5	ခဲ့ ၎ က ပါ တယ်။ | 54920 · တယ် လို့သိ ရ ပါ တယ် | 44239 · ခဲ့ တာ ဖြစ် ပါ တယ် | 36892 · တယ် လို့ ဆို ပါ တယ် | 34060 · တာ ပဲ ဖြစ် ပါ တယ် | 24877

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