

General overview

Corpus	Analytics date	Language
my_1_jsonl.tsv	3/26/2024	Burmese (my)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
239,473	47,772,618	9,899,356 (20.72 %)	501M	8.0 GB	

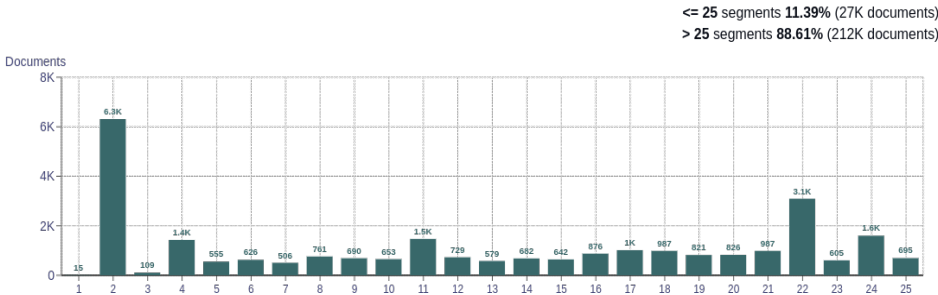
Top 10 domains

Domain	Docs	% of total
blogspot.com	15K	6.19
blogspot.sg	14K	5.71
moemaka.com	13K	5.23
irrawaddy.com	10K	4.23
blogspot.kr	8.1K	3.39
thittoolwin.com	5.9K	2.44
blogspot.ru	3.9K	1.64
mysportmyanmar.com	3.5K	1.47
blogspot.de	3.4K	1.42
ygnnews.com	3K	1.27

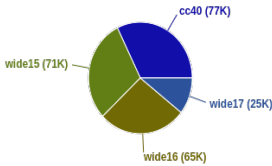
Top 10 TLDs

Domain	Docs	% of total
com	136K	56.72
org	18K	7.68
sg	14K	5.71
net	13K	5.33
com.mm	8.2K	3.42
kr	8.1K	3.39
ru	4K	1.65
xyz	3.6K	1.52
de	3.4K	1.42
in	2.6K	1.09

Documents size (in segments)

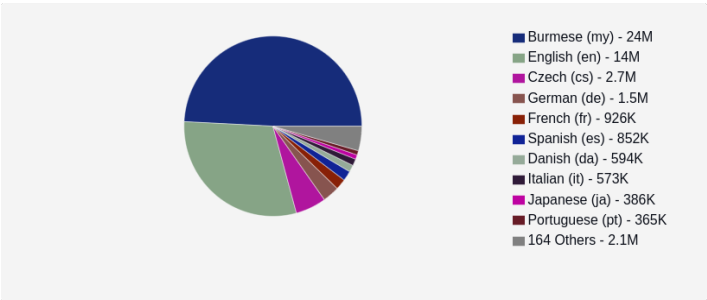


Documents by collection

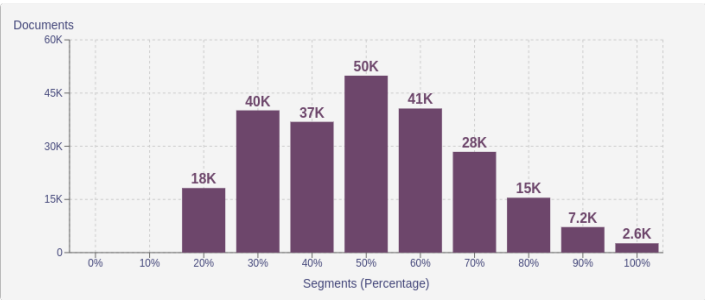


Language Distribution

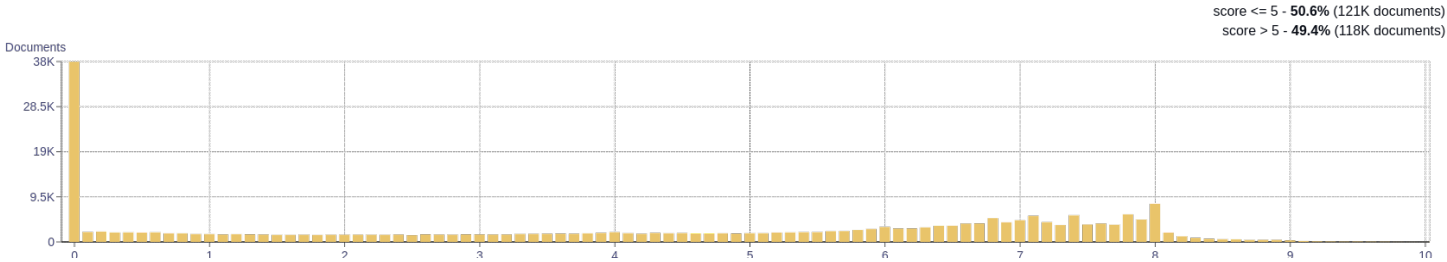
Number of segments



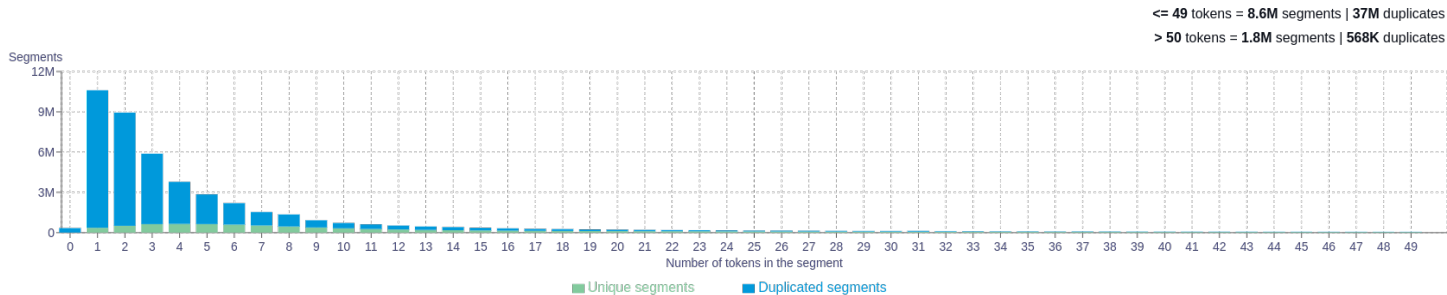
Percentage of segments in Burmese (my) inside documents



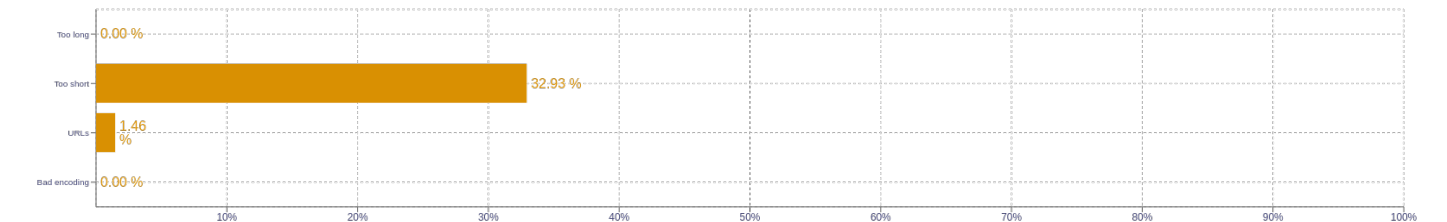
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ကံ့ 8333926 က 8297226 ဝါ 7883766 ခ 3850051 ဖ 3547059
2	ပါ တယူဒါ 1665682 ပါ တယံ 656538 ခု ပါ 540256 ပါ ဘူး 422091 ဆို တာ 415393
3	lsd exception locked 278833 this blog this! 219870 blog this! sharetotwitter 219860 this! sharetotwitter sharetofacebook 219825 email this blog 219644
4	this blog this! sharetotwitter 219846 blog this! sharetotwitter sharetofacebook 219825 email this blog this! 219644 this! sharetotwitter sharetofacebook sharetop 197889 sharetotwitter sharetofacebook sharetop interest 197888
5	this blog this! sharetotwitter sharetofacebook 219811 email this blog this! sharetotwitter 219642 blog this! sharetotwitter sharetofacebook sharetop 197889 this! sharetotwitter sharetofacebook sharetop interest 197888 newer post older post home 41638

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>