HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
bod_Tibt.jsonl.tsv	9/25/2024	Tibetan (bo)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
27,444	464,993	291,415 (62.67 %)	195M	752.19 MB	268,092,556

Top 10 domains

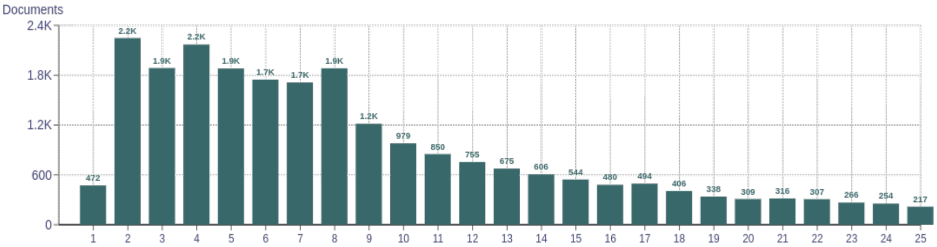
Domain	Docs	% of total
bod.asia	2.8K	10.30
tibettimes.net	1.9K	7.07
tibet3.com	1.6K	5.76
tsadra.org	1.3K	4.89
chithu.org	1.1K	4.09
wikipedia.org	934	3.40
tibet.cn	795	2.90
zangdiyg.com	753	2.74
himalayabon.com	611	2.23
84000.co	587	2.14

Top 10 TLDs

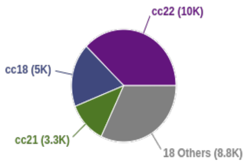
Domain	Docs	% of total
com	10K	37.84
org	6.5K	23.67
asia	2.8K	10.38
net	2.7K	9.78
cn	2.6K	9.32
com.cn	638	2.32
co	587	2.14
gov.cn	409	1.49
ch	200	0.73
us	155	0.56

Documents size (in segments)

<= 25 segments **83.83%** (23K documents)
> 25 segments **16.17%** (4.4K documents)

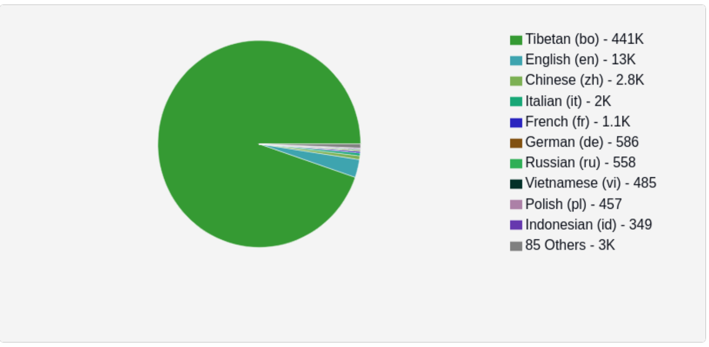


Documents by collection

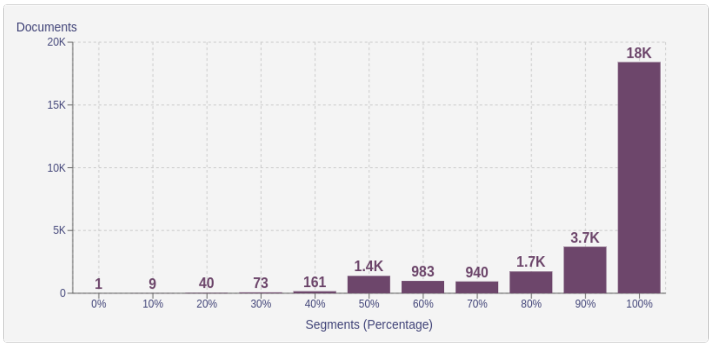


Language Distribution

Number of segments

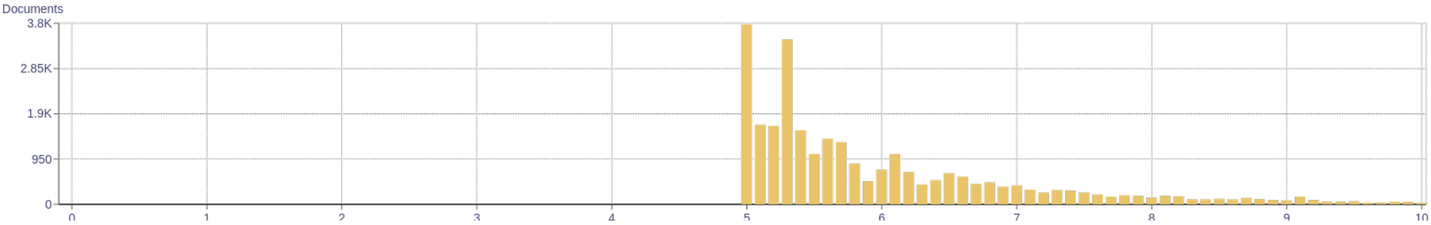


Percentage of segments in Tibetan (bo) inside documents



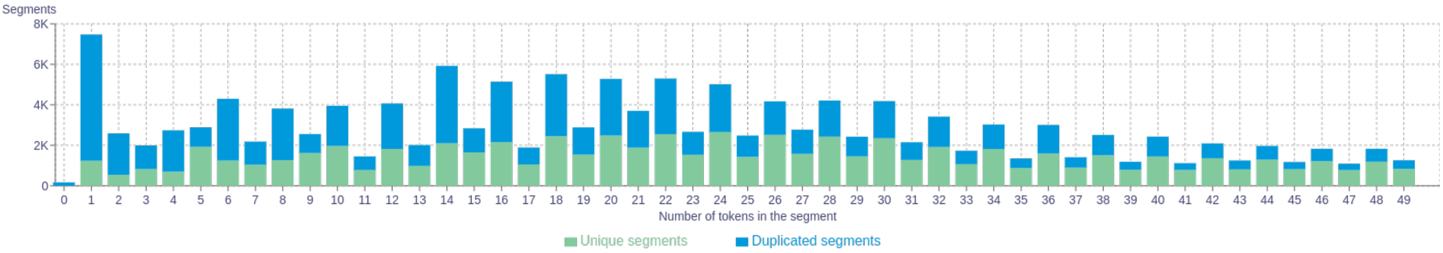
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (27K documents)

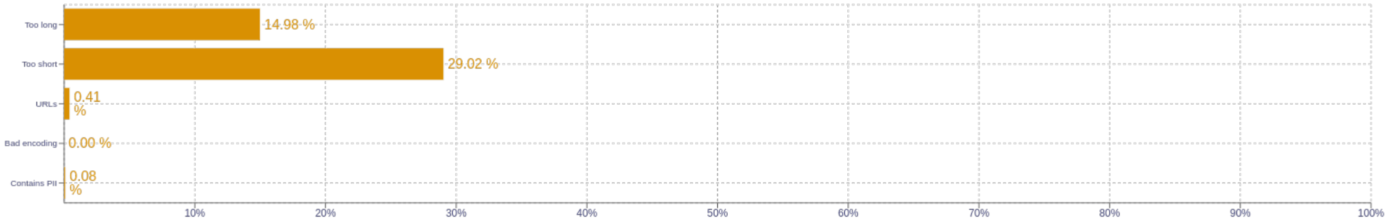


Segment length distribution by token

<= 49 tokens = **72K** segments | **72K** duplicates
> 50 tokens = **321K** segments | **101K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	གས | 2545661 དང | 1558479 བས | 1245383 པར | 1038793 འད | 858898
2	views today | 1426 total views | 1426 of the | 986 posted by | 962 folios 1a1 | 953
3	folios 1a1 to | 953 the rest of | 852 rest of this | 852 read the rest | 852 of this entry | 852
4	the rest of this | 852 rest of this entry | 852 read the rest of | 852 jamgön kongtrul lodrö taye | 148 terdak lingpa gyurme dorje | 109
5	the rest of this entry | 852 read the rest of this | 852 rdzogs pa chen po sde | 74 pa chen po sde gsum | 74 dam chos rdzogs pa chen | 74

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