HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
dzo_Tibt.jsonl.tsv	10/30/2024	Dzongkha (dz)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,626	39,971	18,450 (46.16 %)	5M	19.8 MB	7,336,913

Top 10 domains

Domain	Docs	% of total
blogspot.com	477	29.34
dzkuensel.com	429	26.38
libreoffice.org	153	9.41
ecb.bt	83	5.10
gnhc.gov.bt	50	3.08
wikipedia.org	49	3.01
investigative-manual.org	45	2.77
virginia.edu	37	2.28
library.gov.bt	31	1.91
dzkuensel.bt	28	1.72

Top 10 TLDs

Domain	Docs	% of total
com	938	57.69
org	251	15.44
gov.bt	212	13.04
bt	138	8.49
edu	37	2.28
edu.bt	12	0.74
in	9	0.55
net	8	0.49
nl	7	0.43
de	7	0.43

## Documents size (in segments)

<= 25 segments **87.82%** (1.4K documents)
**> 25** segments **12.18%** (196 documents)
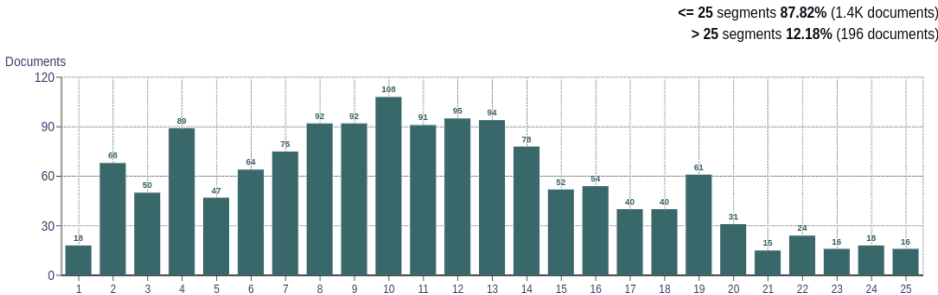
Documents by collection

cc22 (658)
cc18 (257)
cc17 (231)
cc21 (225)
16 Others (255)
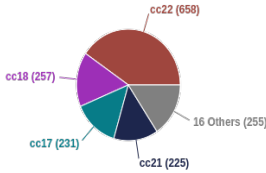
Language Distribution

Number of segments

- Tibetan (bo) - 38K
- English (en) - 1.4K
- Italian (it) - 130
- Indonesian (id) - 99
- Russian (ru) - 72
- German (de) - 23
- Hindi (hi) - 18
- French (fr) - 17
- Turkish (tr) - 12
- Chinese (zh) - 11
- 36 Others - 119

*Dzongkha (dz) identification might be inaccurate because language is not supported by Fasttext
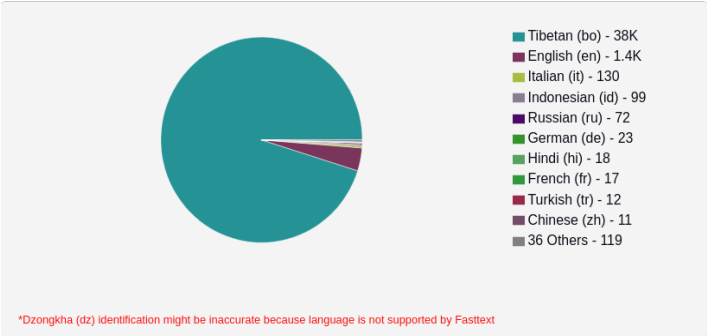
Percentage of segments in Dzongkha (dz) inside documents
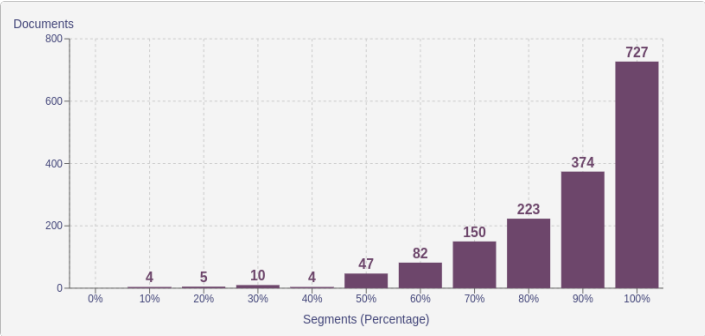
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (1.6K documents)

Segment length distribution by token

<= 49 tokens = **4.7K** segments | **10K** duplicates
> 50 tokens = **25K** segments | **11K** duplicates

Number of tokens in the segment

- Unique segments
- Duplicated segments

Segment noise distribution

Too long	0.00 %
Too short	15.49 %
URLs	0.22 %
Bad encoding	0.00 %
Contains PII	0.06 %
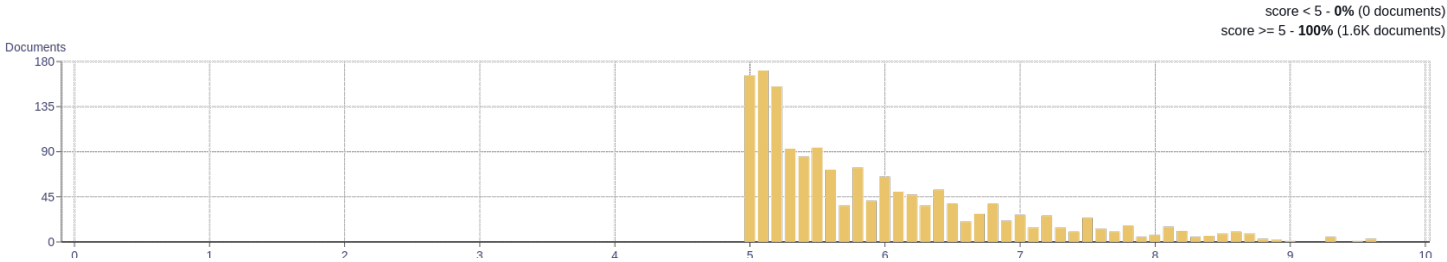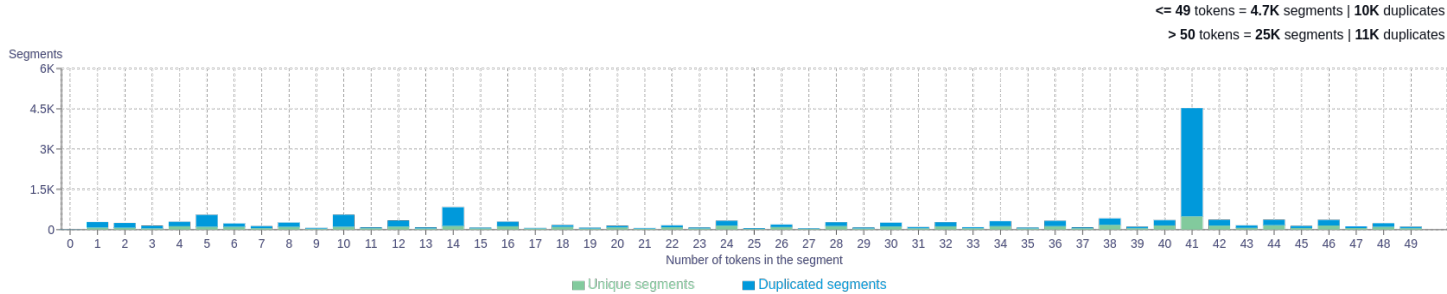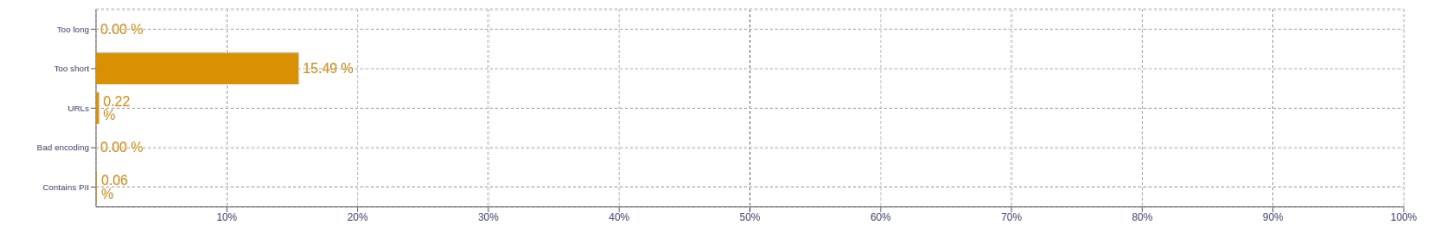
About HPLT Analytics

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