HPLT Analytics report

General overview

Corpus	Analytics date	Language
tum_Latn.jsonl.tsv	11/7/2024	Tumbuka (tum)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
4,384	99,013	65,730 (66.39 %)	3.6M	20.52 MB	21,003,570

Top 10 domains

Domain	Docs	% of total
jw.org	4K	92.02
wikipedia.org	174	3.97
globalrecordings.net	40	0.91
nkhanimchitumbuka.com	31	0.71
gpedia.com	11	0.25
bible.is	10	0.23
watchtower.org	10	0.23
digitallibrary.io	8	0.18
bibleforchildren.org	5	0.11
sciencegraph.net	5	0.11

Top 10 TLDs

Domain	Docs	% of total
org	4.2K	96.58
com	69	1.57
net	48	1.09
io	11	0.25
is	10	0.23
info	5	0.11
co.uk	4	0.09
click	1	0.02
no	1	0.02
vn	1	0.02

Documents size (in segments)

<= 25 segments **70.96%** (3.1K documents)
> 25 segments **29.04%** (1.3K documents)
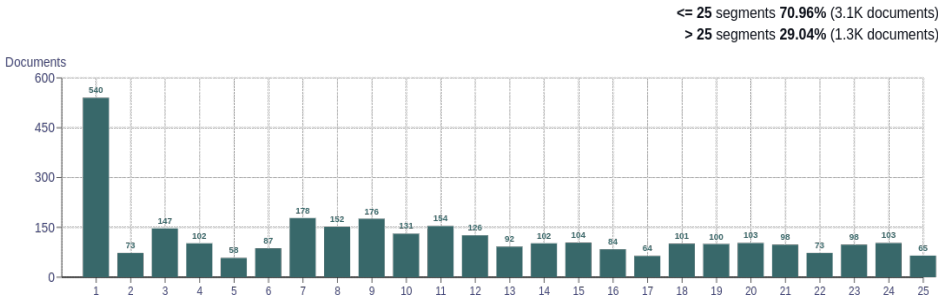
Documents by collection

cc17 (1.8K)
cc18 (844)
wide16 (679)
18 Others (1K)
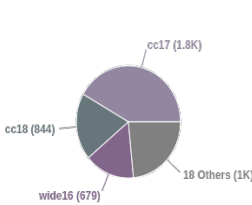
Language Distribution

Number of segments

- English (en) - 34K
- Swahili (sw) - 19K
- Filipino (tl) - 10K
- Indonesian (id) - 9.5K
- Spanish (es) - 4.7K
- Italian (it) - 3.5K
- Esperanto (eo) - 2.2K
- Croatian (hr) - 2K
- Polish (pl) - 2K
- Waray (war) - 1.7K
- 103 Others - 11K

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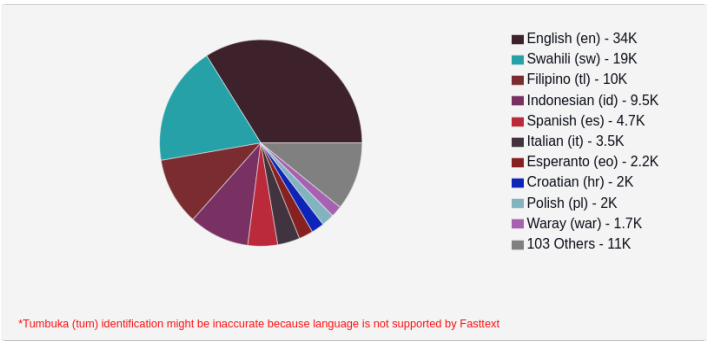
Percentage of segments in Tumbuka (tum) inside documents
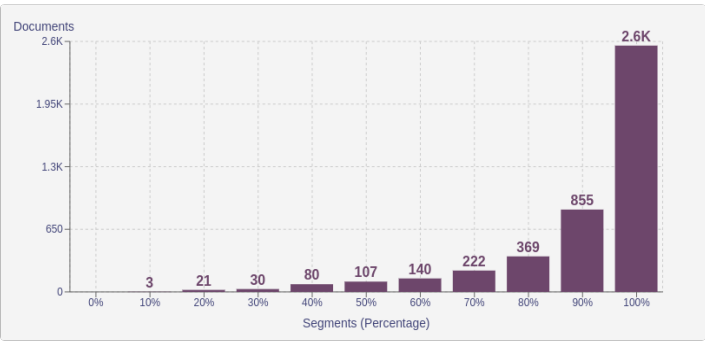
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (4.4K documents)

Segment length distribution by token

<= 49 tokens = **48K** segments | **28K** duplicates
> 50 tokens = **23K** segments | **5K** duplicates

Number of tokens in the segment

- Unique segments
- Duplicated segments

Segment noise distribution

- Too long: 0.00 %
- Too short: 11.55 %
- URLs: 0.98 %
- Bad encoding: 0.00 %
- Contains PII: 0.00 %
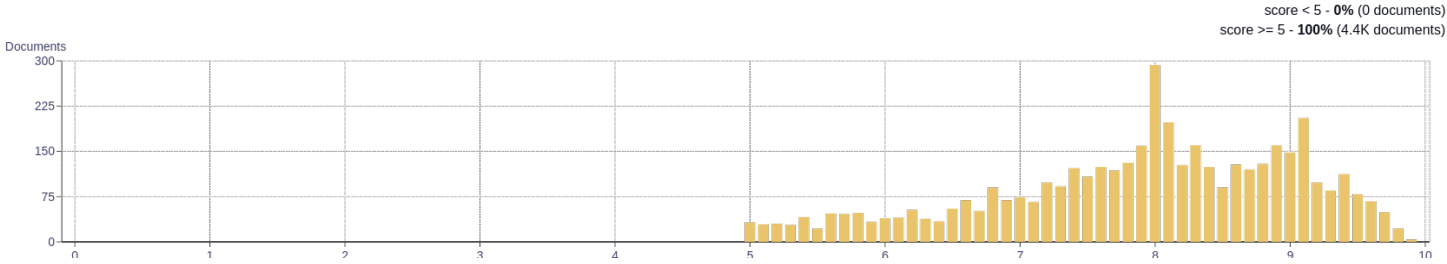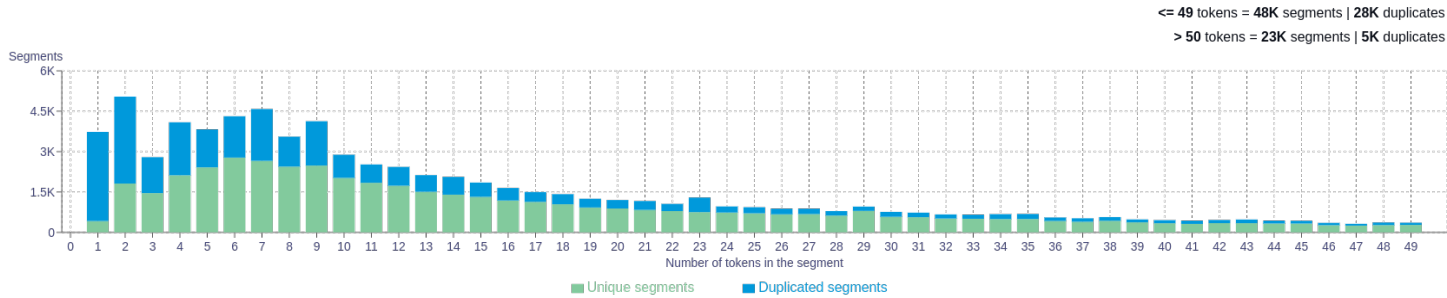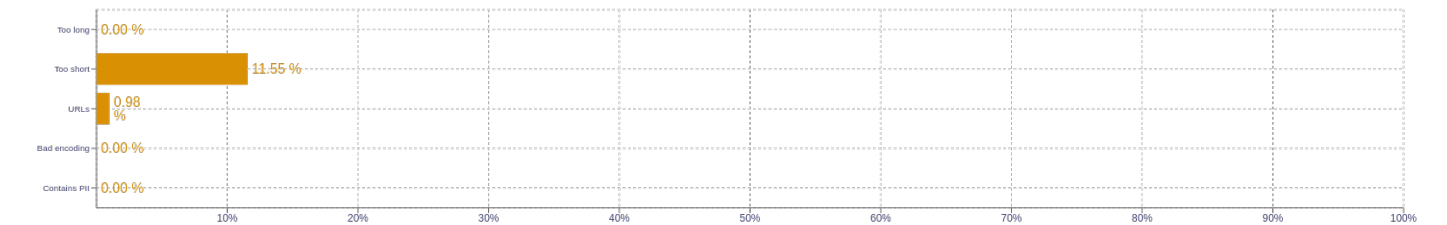
About HPLT Analytics

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