HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
umb_Latn.jsonl.tsv	11/4/2024	Umbundu (umb)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
2,471	59,912	41,987 (70.08 %)	2.9M	14.8 MB	15,348,339

Top 10 domains

Domain	Docs	% of total
jw.org	1.9K	78.19
neweralive.na	167	6.76
bible.is	108	4.37
kundana.com.na	95	3.84
blogspot.com	28	1.13
globalrecordings.net	26	1.05
pngscriptures.org	19	0.77
watchtower.org	12	0.49
unicode.org	11	0.45
blogspot.pt	7	0.28

Top 10 TLDs

Domain	Docs	% of total
org	2K	80.49
na	167	6.76
is	108	4.37
com.na	98	3.97
com	57	2.31
net	29	1.17
pt	7	0.28
bible	4	0.16
co	3	0.12
cc	3	0.12

Documents size (in segments)

<= 25 segments **68.68%** (1.7K documents)
> 25 segments **31.32%** (773 documents)
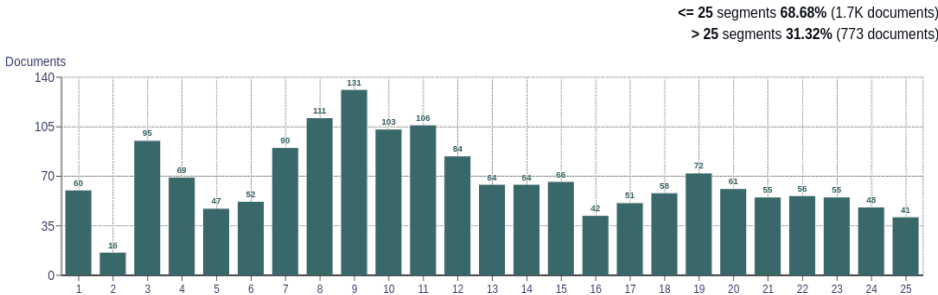
Documents by collection

cc17 (594), wide16 (653), wide12 (253), 17 Others (971)
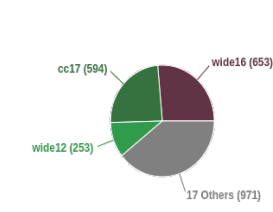
Language Distribution

Number of segments

- English (en) - 14K
- Esperanto (eo) - 9.3K
- Filipino (tl) - 4.5K
- Swahili (sw) - 4.4K
- Indonesian (id) - 3.5K
- Spanish (es) - 3.1K
- Italian (it) - 2.8K
- Croatian (hr) - 2.1K
- Turkish (tr) - 2.1K
- Finnish (fi) - 2.1K
- 108 Others - 12K

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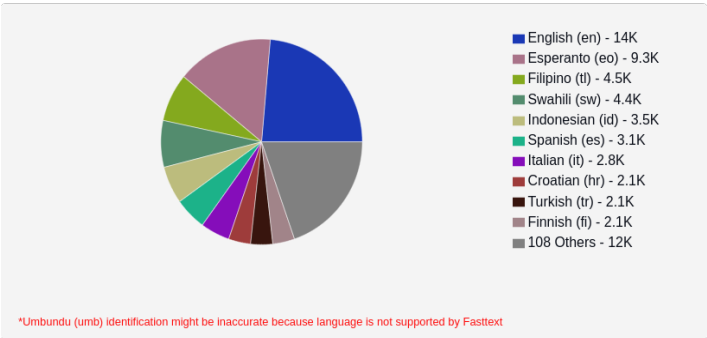
Percentage of segments in Umbundu (umb) inside documents
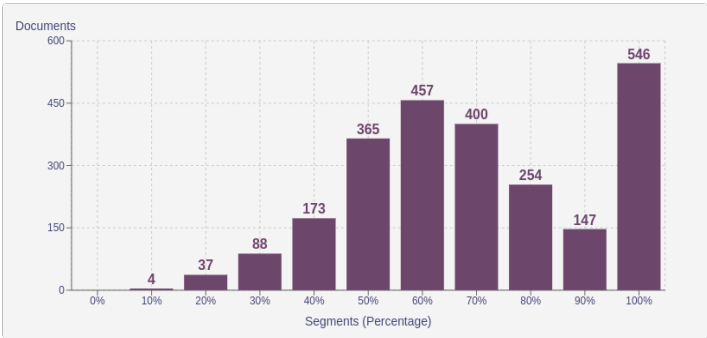
## Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (2.5K documents)

Segment length distribution by token

<= 49 tokens = 26K segments | **13K** duplicates
> 50 tokens = 21K segments | **5.3K** duplicates

Number of tokens in the segment

- Unique segments
- Duplicated segments

Segment noise distribution

Too long	0.00 %
Too short	5.01 %
URLs	0.21 %
Bad encoding	0.31 %
Contains PII	0.05 %
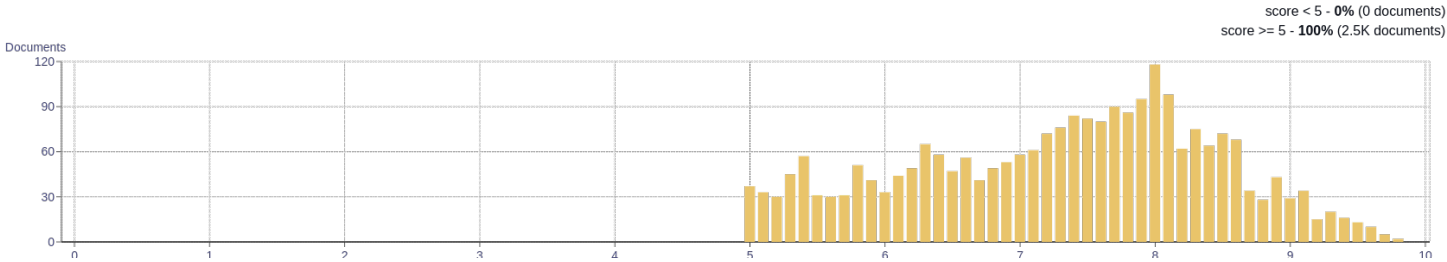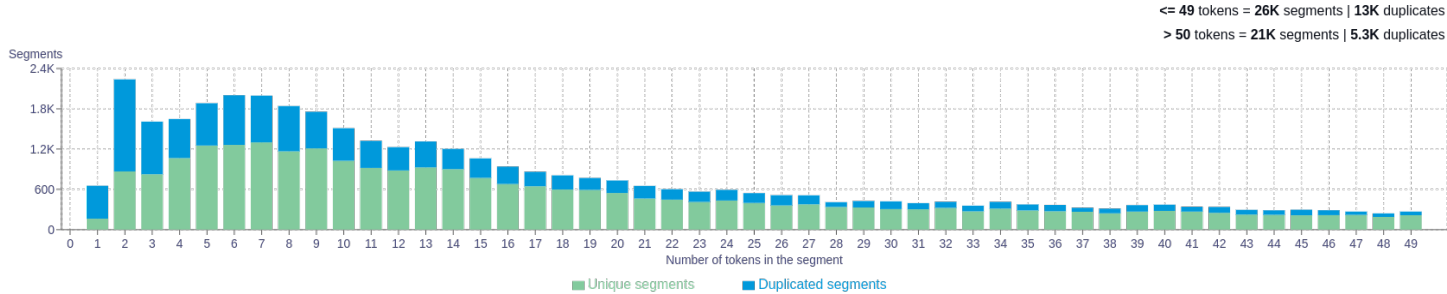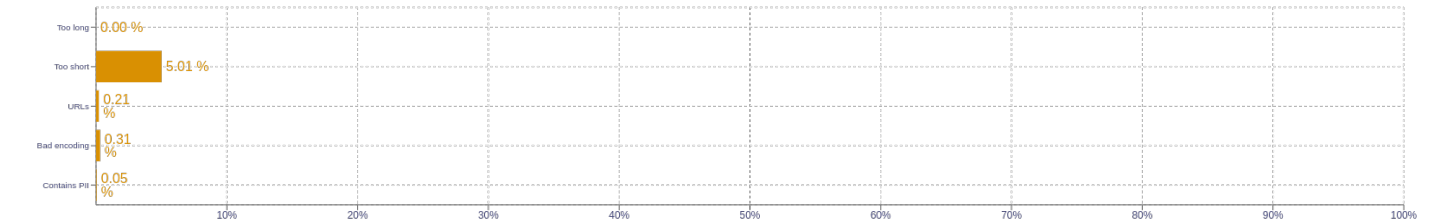
About HPLT Analytics

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