HPLT Analytics report

HPLT Analytics

General overview

Corpus	Analytics date	Language
kmb_Latn.jsonl.tsv	11/4/2024	Kimbundu (kmb)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
531	11,799	7,671 (65.01 %)	476K	2.0 MB	2,056,854

Top 10 domains

Domain	Docs	% of total
jw.org	423	79.66
zedge.net	13	2.45
nudeindianbhabhiclub.com	9	1.69
unicode.org	9	1.69
dramasonline.pk	8	1.51
reunion.com	7	1.32
contafrica.org	6	1.13
ohchr.org	5	0.94
redwap.vip	4	0.75
watchtower.org	4	0.75

Top 10 TLDs

Domain	Docs	% of total
org	451	84.93
com	37	6.97
net	15	2.82
pk	8	1.51
ao	4	0.75
vip	4	0.75
co	3	0.56
de	2	0.38
mobi	1	0.19
club	1	0.19

Documents size (in segments)

<= 25 segments **74.01%** (393 documents)
> 25 segments **25.99%** (137 documents)
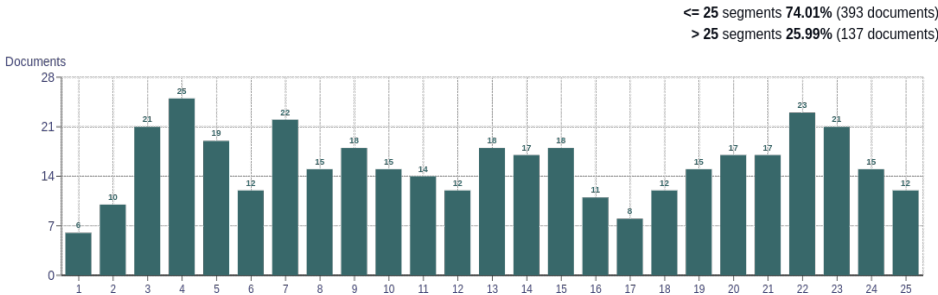
Documents by collection

wide16 (207)
wide12 (91)
wide11 (72)
16 Others (161)
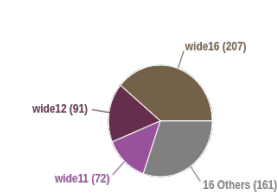
Language Distribution

Number of segments

- Spanish (es) - 2.7K
- English (en) - 2K
- Swahili (sw) - 707
- Indonesian (id) - 552
- Finnish (fi) - 495
- Filipino (tl) - 487
- Esperanto (eo) - 477
- Sundanese (su) - 405
- Slovenian (sl) - 356
- Polish (pl) - 289
- 84 Others - 3.3K

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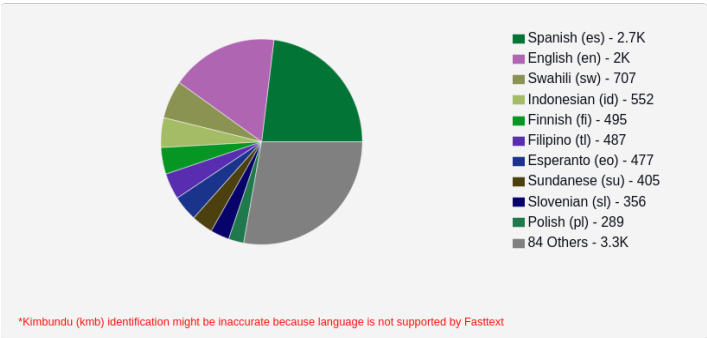
Percentage of segments in Kimbundu (kmb) inside documents
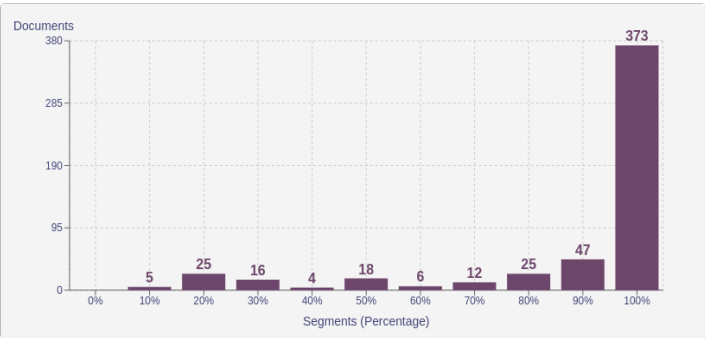
## Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (531 documents)

Segment length distribution by token

<= 49 tokens = **5.3K** segments | **3.2K** duplicates
> 50 tokens = **3.3K** segments | **885** duplicates

- Unique segments
- Duplicated segments

Segment noise distribution

- Too long: 0.00 %
- Too short: 6.37 %
- URLs: 0.69 %
- Bad encoding: 0.39 %
- Contains PII: 0.00 %
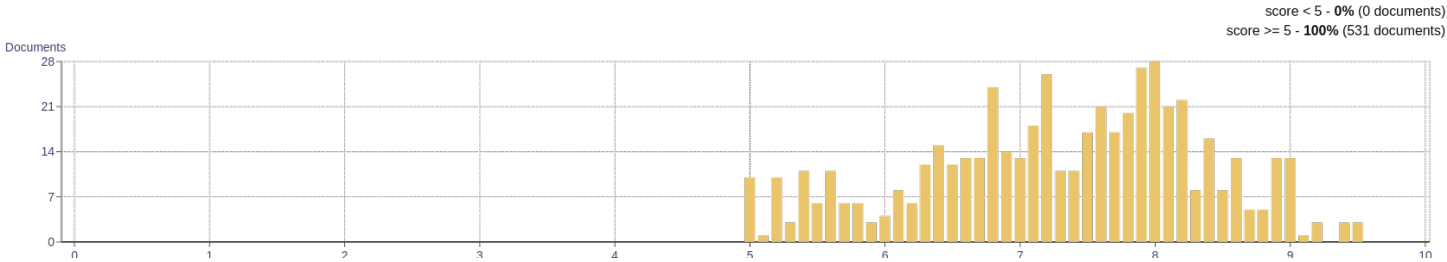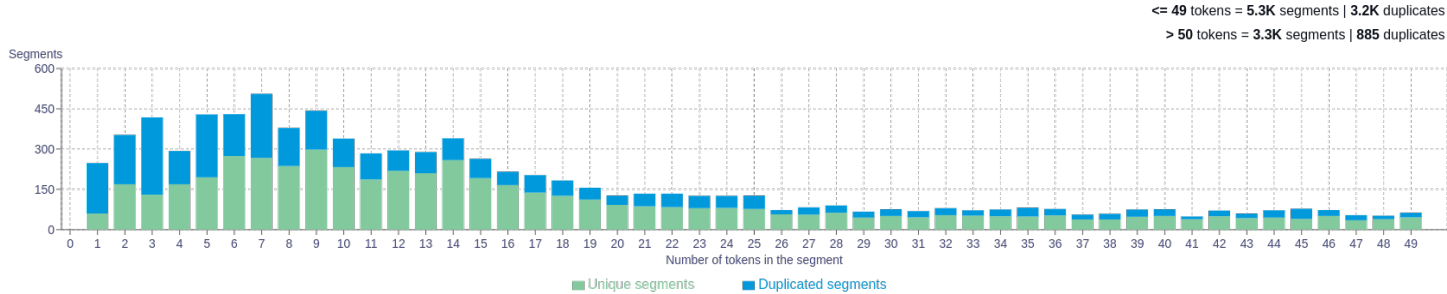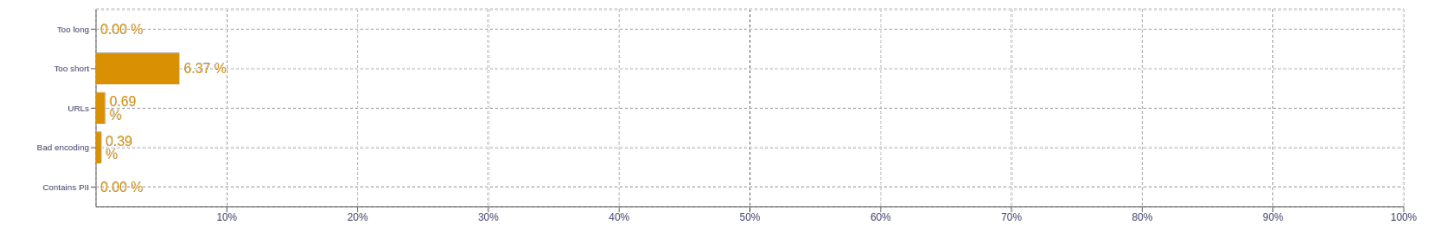
About HPLT Analytics

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