HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
kam_Latn.jsonl.tsv	12/5/2024	Kamba (kam)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,183	14,259	9,563 (67.07 %)	867K	4.9 MB	4,631,764

Top 10 domains

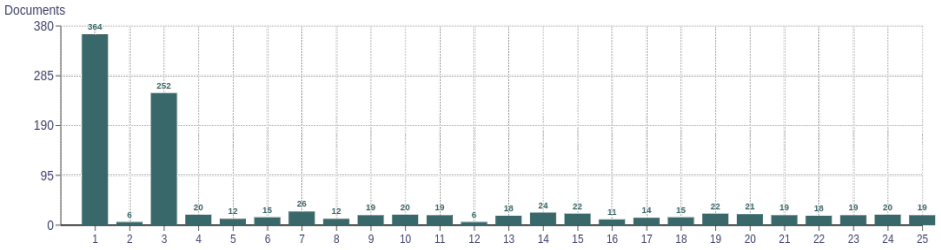
Domain	Docs	% of total
bible.is	583	49.28
jw.org	510	43.11
sangufm.co.ke	32	2.70
kituionline.com	9	0.76
gospelgo.com	6	0.51
jaladaafrica.org	5	0.42
watchtower.org	5	0.42
graduates.com	4	0.34
rmsradio.co.ke	4	0.34
4laws.com	3	0.25

Top 10 TLDs

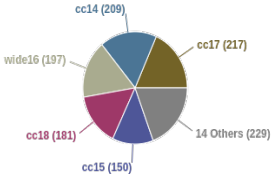
Domain	Docs	% of total
is	583	49.28
org	526	44.46
co.ke	38	3.21
com	31	2.62
net	3	0.25
io	1	0.08
web.id	1	0.08

Documents size (in segments)

<= 25 segments **85.63%** (1K documents)
> 25 segments **14.37%** (170 documents)

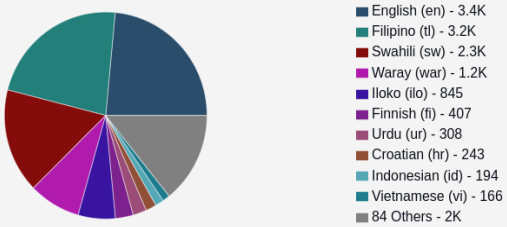


Documents by collection



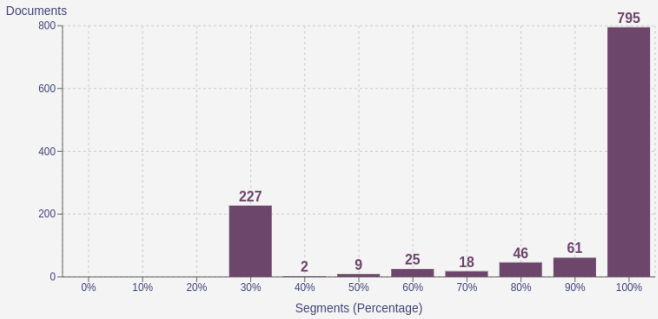
Language Distribution

Number of segments



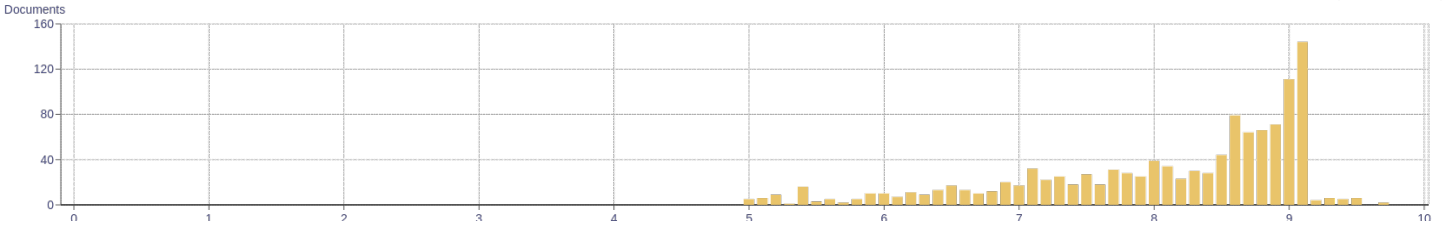
*Kamba (kam) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Kamba (kam) inside documents



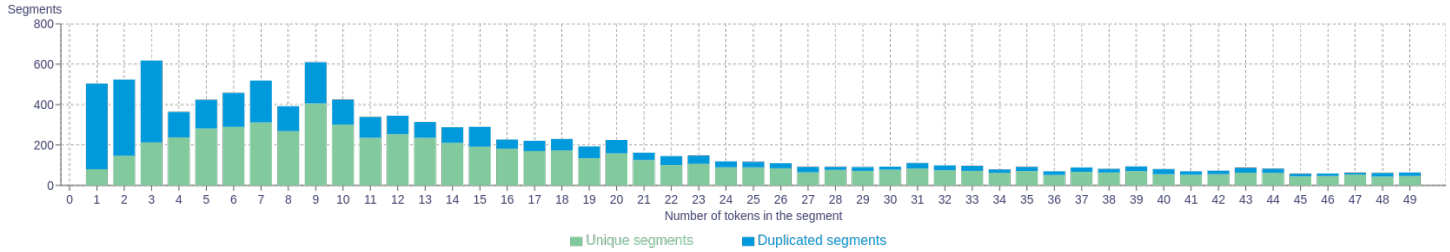
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (1.2K documents)

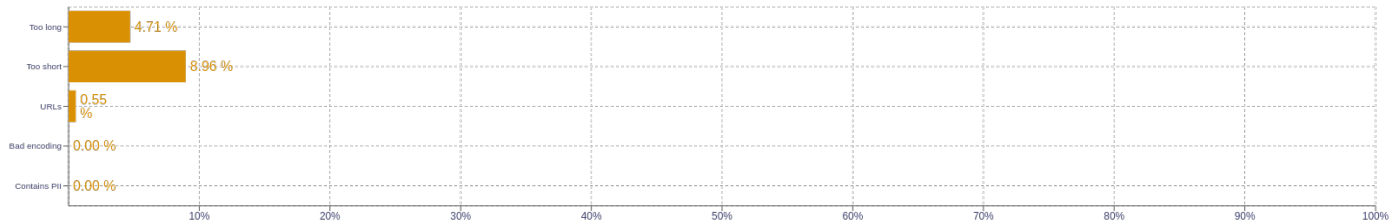


Segment length distribution by token

<= 49 tokens = **6.5K** segments | **3.7K** duplicates
> 50 tokens = **4.1K** segments | **1K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	kwa | 4012 · alĩ | 3916 · andũ | 3849 · yeova | 3785 · sika | 3780
2	pĩna alĩ | 424 · sya yeova | 408 · ngũsĩ sya | 397 · kũũ nthĩ | 372 · sika alĩ | 362
3	ngũsĩ sya yeova | 393 · ĩũlũ wa nthĩ | 320 · tene na tene | 261 · atamu na eva | 181 · munumba ya kĩnzua | 177
4	kĩshekulu wĩĩtu yesu kilisito | 93 · maũndũ ma vata kuma | 56 · asu asu asu asu | 49 · ũu wĩ o vo | 46 · ũvoo mũseo wa ũsumbĩ | 42
5	maũndũ ma vata kuma ndetonĩ | 53 · asu asu asu asu asu | 36 · ĩndĩ o na ũu wĩ | 35 · soma maelesyo ma mũthya namba | 35 · ũsumbĩ wa ngai nĩ kyaũ | 34

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