

General overview

Corpus	Analytics date	Language
tum_Latn.jsonl.tsv	11/27/2024	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

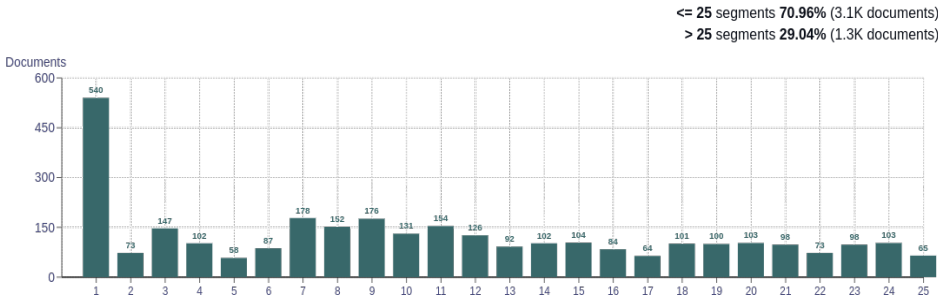
Top 10 domains

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
digitalibrary.io	8	0.18
bibleforchildren.org	5	0.11
sciencegraph.net	5	0.11

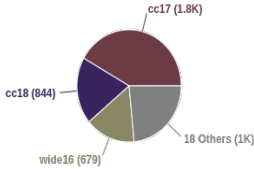
Top 10 TLDs

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)

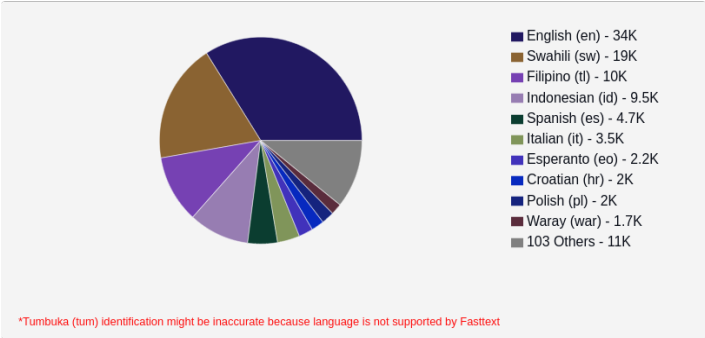


Documents by collection

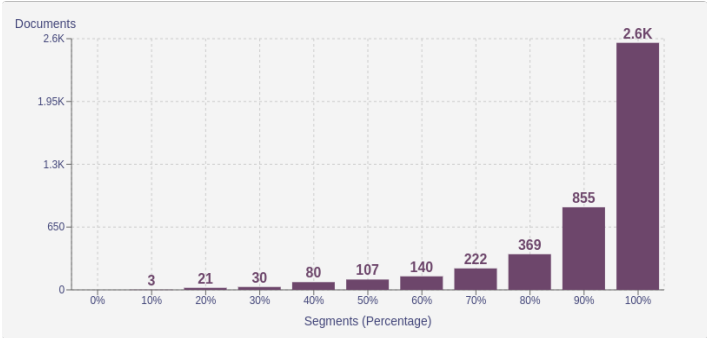


Language Distribution

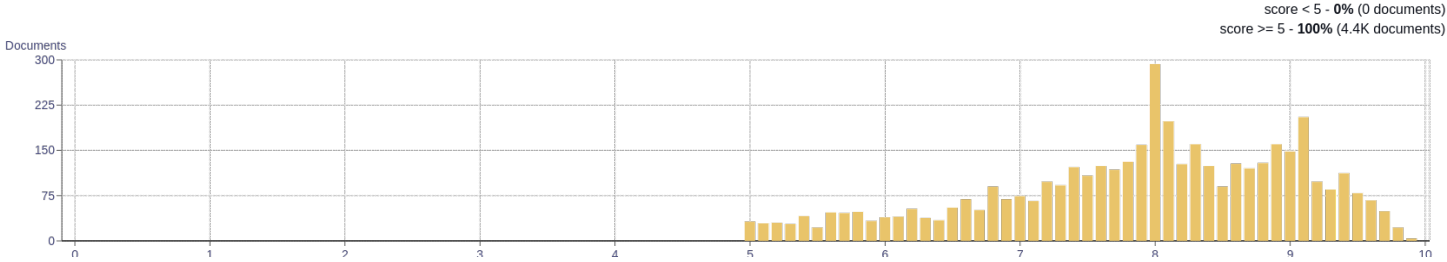
Number of segments



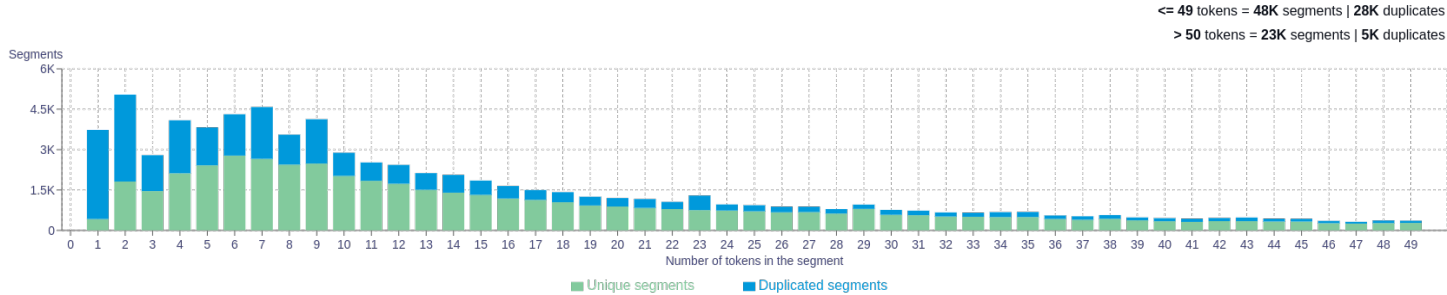
Percentage of segments in Tumbuka (tum) inside documents



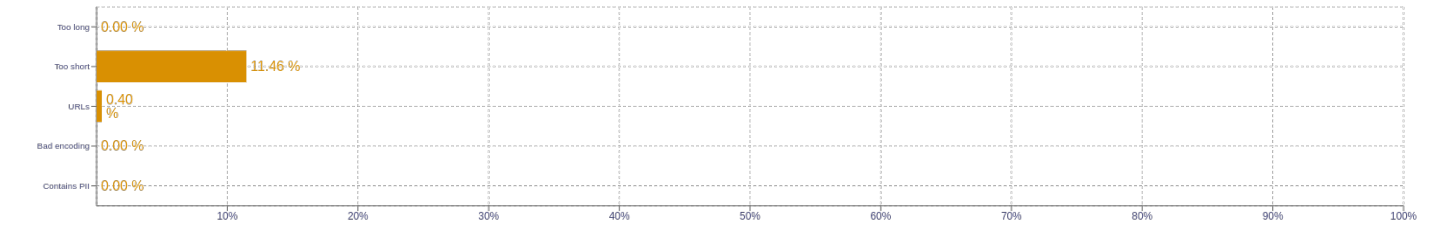
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	yehova 27729 ŵanthu 24888 chiuta 23684 ndi 21388 yesu 20680
2	ŵa yehova 2263 charu chapasi 1962 ŵakaboni ŵa 1822 nga umo 1620 ŵanthu ŵanandi 1416
3	ŵakaboni ŵa yehova 1804 ufumu wa chiuta 877 anamuliro gha charu 854 mazgu gha chiuta 731 akaboni aku yehova 686
4	anamuliro gha charu chiphya 854 malemba ghakupatulika mu mang 681 ŵazgani baibolo pa intaneti 532 lero lero lero lero 490 baibolo la pa intaneti 317
5	lero lero lero lero lero 482 ŵanthu ŵanthu ŵanthu ŵanthu ŵanthu 98 ndicho fumu yikuru yehova yayowoya 89 kumbi ndi vinthu wuli vo 83 sambiro la baibolo la mpingo 65

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of types (uniques)/number of tokens", after removing punctuation (<https://www.stinfo.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>, , , 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>