

General overview

Corpus	Analytics date	Language
kmb_Latn.jsonl.tsv	12/3/2024	Kimbundu (kmb)

Volumes

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

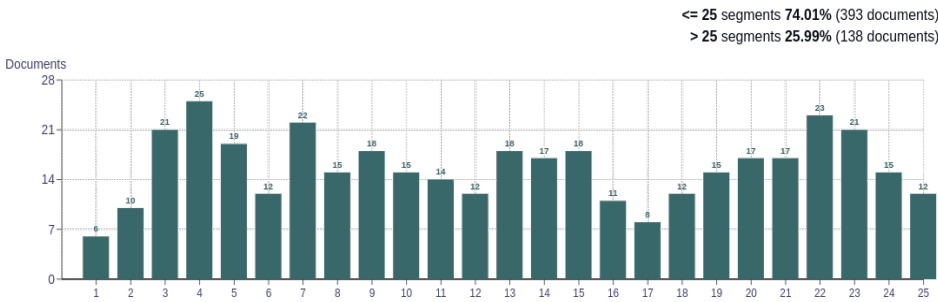
Top 10 domains

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

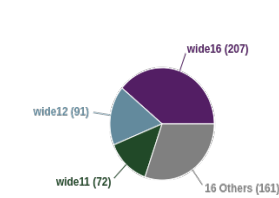
Top 10 TLDs

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)

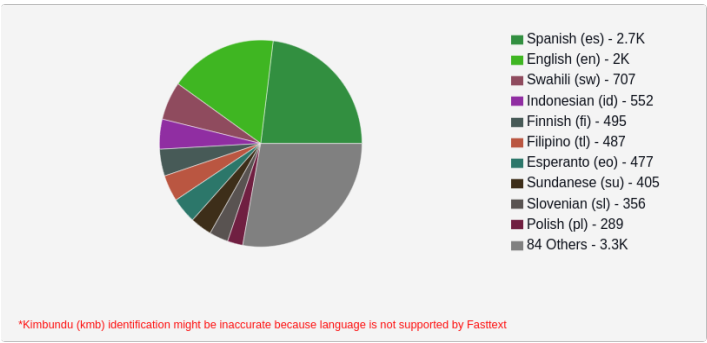


Documents by collection

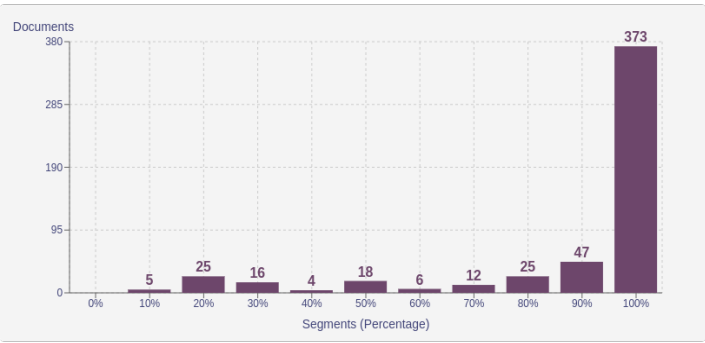


Language Distribution

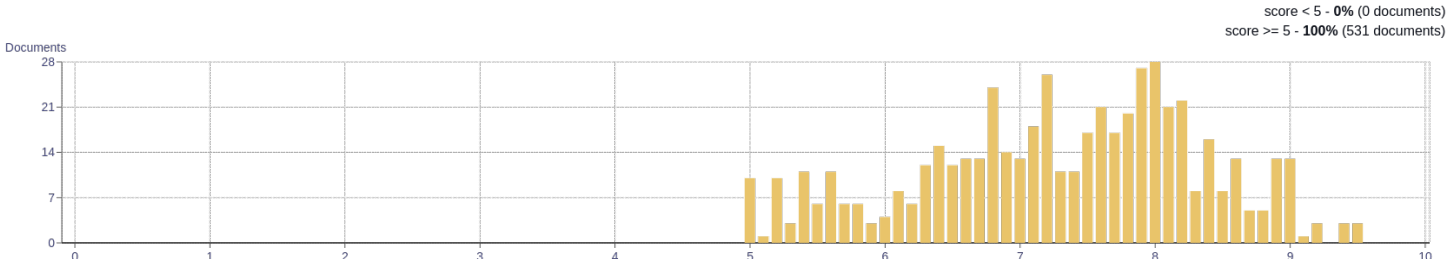
Number of segments



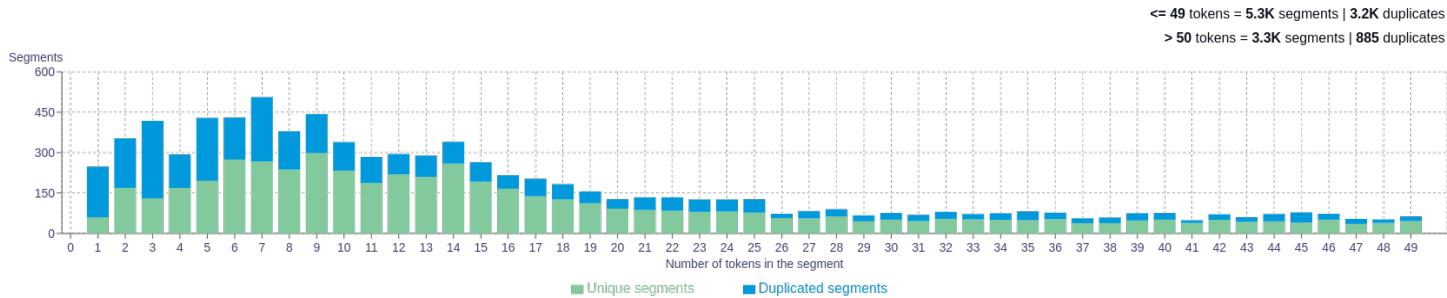
Percentage of segments in Kimbundu (kmb) inside documents



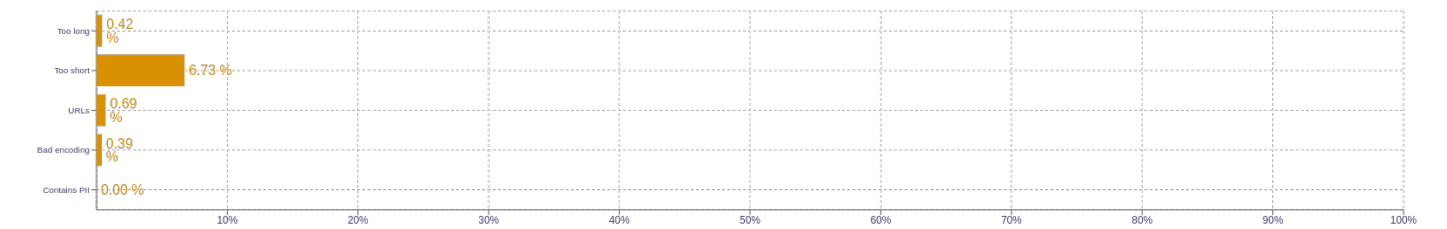
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div><div>o 22093</div><div>mu 15907</div><div>ni 11946</div><div>ua 9341</div><div>ia 8228</div></div>
2	<div><div>o bibidia 1200</div><div>lungu ni 1016</div><div>mukonda díahi 910</div><div>muéne ua 842</div><div>ia lungu 782</div></div>
3	<div><div>ia lungu ni 764</div><div>utuminu ua nzambi 333</div><div>o bibidia i 307</div><div>longa o bibidia 284</div><div>jimbangi ja jihova 266</div></div>
4	<div><div>o utuminu ua nzambi 222</div><div>kia moring kia moring 160</div><div>moring kia moring kia 158</div><div>ixi ixi ixi ixi 151</div><div>o vondadi ia nzambi 147</div></div>
5	<div><div>moring kia moring kia moring 158</div><div>kia moring kia moring kia 158</div><div>ixi ixi ixi ixi ixi 150</div><div>ihí i longa o bibidia 139</div><div>mutu uoso uoso uala ni 90</div></div>

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.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>, , , 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>