

General overview

Corpus	Analytics date	Language
kik_Latn.jsonl.tsv	12/5/2024	Kikuyu (ki)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
3,995	51,931	33,013 (63.57 %)	1.8M	9.92 MB	9,244,068

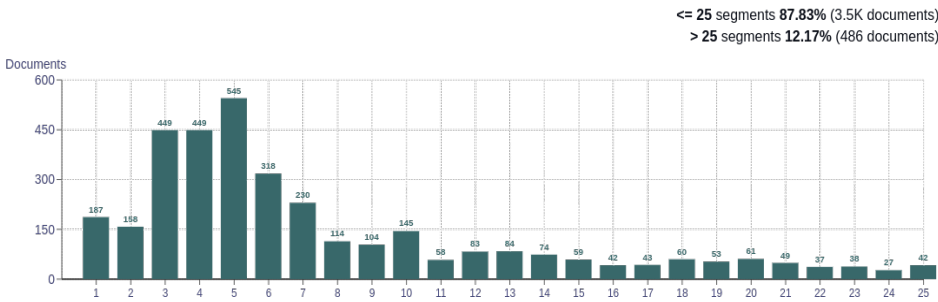
Top 10 domains

Domain	Docs	% of total
mediamaxnetwork.co.ke	1.6K	40.48
jw.org	713	17.85
wikipedia.org	355	8.89
bible.is	224	5.61
biblica.com	191	4.78
blogspot.jp	91	2.28
thekenyatoday.com	69	1.73
blogspot.com	66	1.65
radiokimuri.co.ke	64	1.60
blogspot.fr	38	0.95

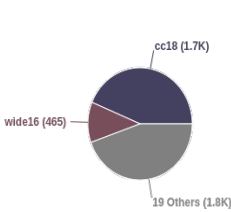
Top 10 TLDs

Domain	Docs	% of total
co.ke	1.9K	47.53
org	1.2K	28.96
com	540	13.52
is	224	5.61
jp	91	2.28
fr	43	1.08
de	15	0.38
net	13	0.33
eu	3	0.08
click	1	0.03

Documents size (in segments)

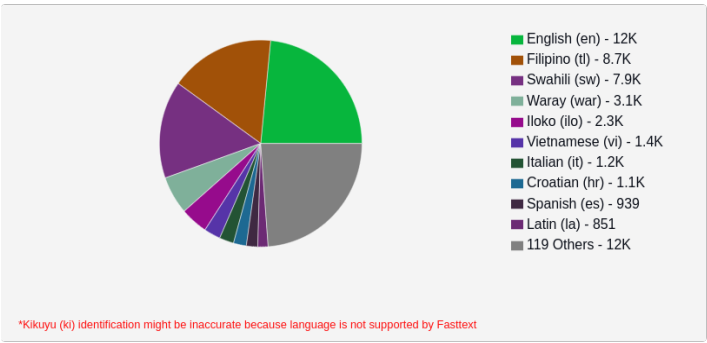


Documents by collection

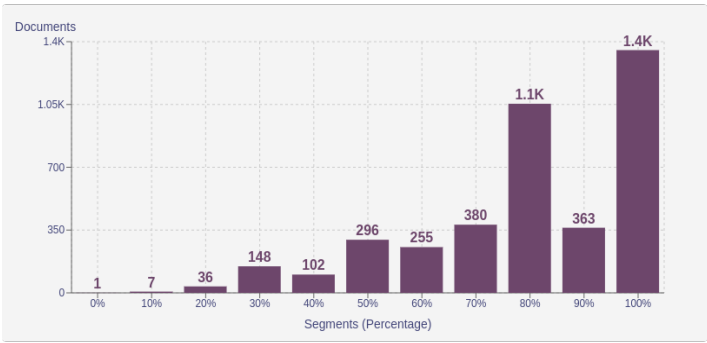


Language Distribution

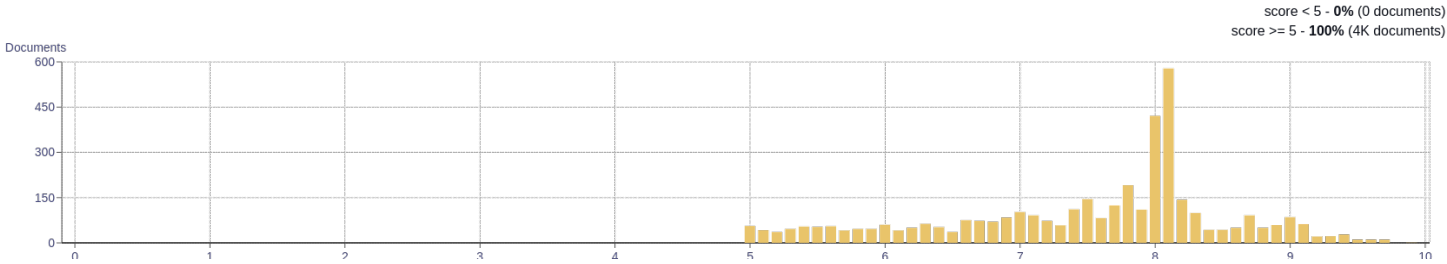
Number of segments



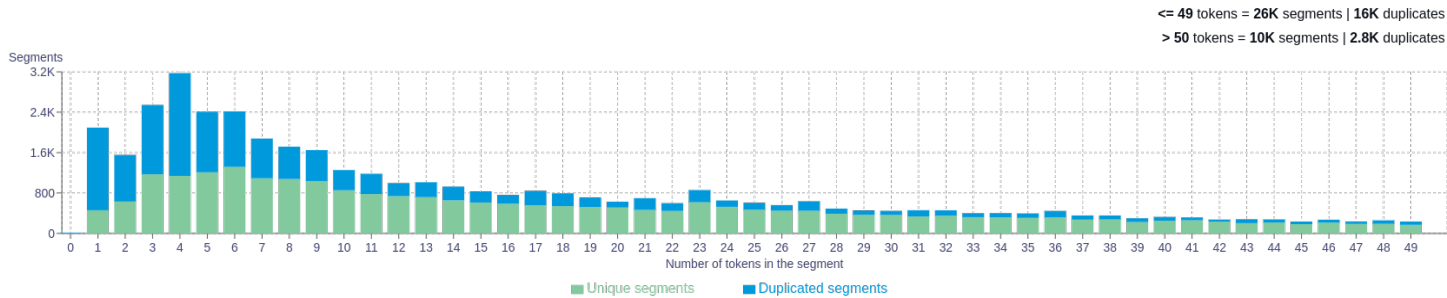
Percentage of segments in Kikuyu (ki) inside documents



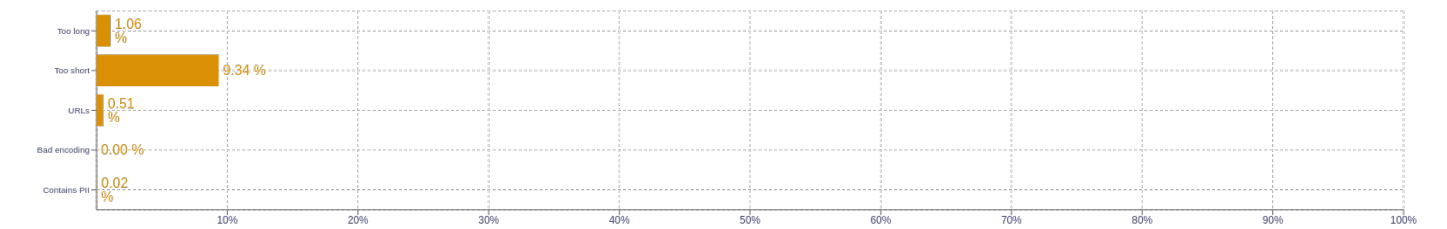
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div><div>wa 44242</div><div>cia 9859</div><div>ngai 9073</div><div>inĩ 7379</div><div>andũ 7227</div></div>
2	<div><div>thĩini wa 2380</div><div>ũndũ wa 1827</div><div>thutha wa 1790</div><div>wa ngai 1457</div><div>ũndũ ũcio 1396</div></div>
3	<div><div>kameme fmkameme fm 1146</div><div>fmkameme fm news 979</div><div>mutongoria wa bururi 521</div><div>tene na tene 492</div><div>aira a jehova 405</div></div>
4	<div><div>kameme fmkameme fm news 979</div><div>mutongoria wa bururi uhuru 295</div><div>wa bururi uhuru kenyatta 261</div><div>wa tene na tene 202</div><div>kameme fmkameme fm newskameme 164</div></div>
5	<div><div>mutongoria wa bururi uhuru kenyatta 257</div><div>kameme fmkameme fm newskameme news 162</div><div>muoyo wa tene na tene 134</div><div>ariũ na aarĩ a ithe 87</div><div>first on mediamax network limited 85</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>