HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
sw_1.jsonl.tsv	3/20/2024	Swahili (sw)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
698,565	76,253,152	17,500,605 (22.95 %)	862M	4.09 GB	

Top 10 domains

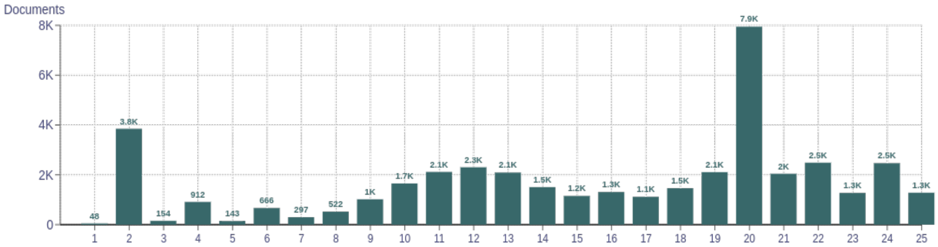
Domain	Docs	% of total
diebuchsuche.com	407K	58.24
fanpop.com	18K	2.56
freelancer.co.ke	16K	2.28
tuko.co.ke	9.7K	1.39
blogspot.com	8.9K	1.27
mwanahalisionline.com	6.9K	0.99
airbnb.com	6.2K	0.89
blogspot.co.uk	5.5K	0.79
w3eacademy.com	5K	0.72
teknokona.com	4.6K	0.66

Top 10 TLDs

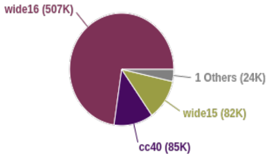
Domain	Docs	% of total
com	539K	77.18
co.ke	33K	4.77
co.tz	31K	4.49
org	19K	2.74
go.tz	9.3K	1.33
co.uk	5.5K	0.79
fr	4.7K	0.68
nl	3.4K	0.48
se	3.2K	0.45
no	3.1K	0.45

Documents size (in segments)

<= 25 segments **6%** (42K documents)
> 25 segments **94%** (657K documents)

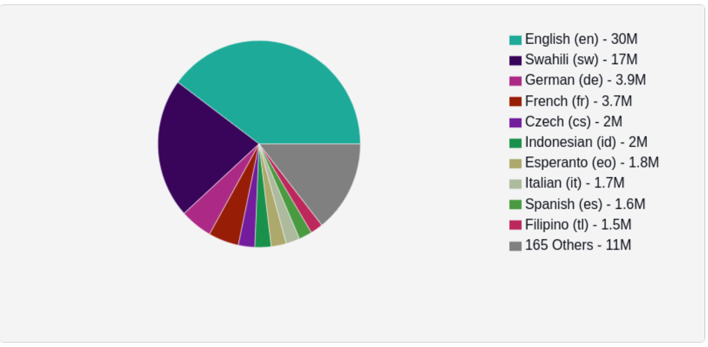


Documents by collection

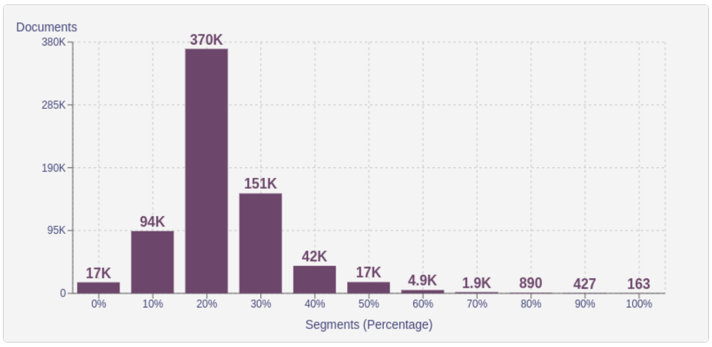


Language Distribution

Number of segments

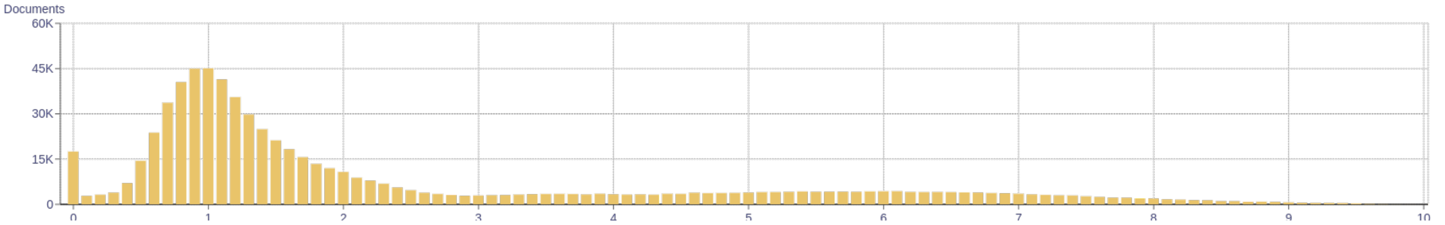


Percentage of segments in Swahili (sw) inside documents



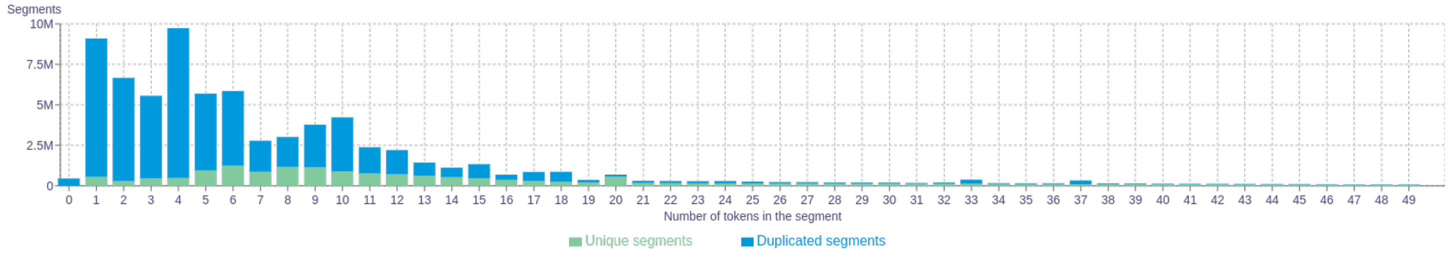
Distribution of documents by document score

score < 5 - **82.19%** (574K documents)
score >= 5 - **17.81%** (124K documents)

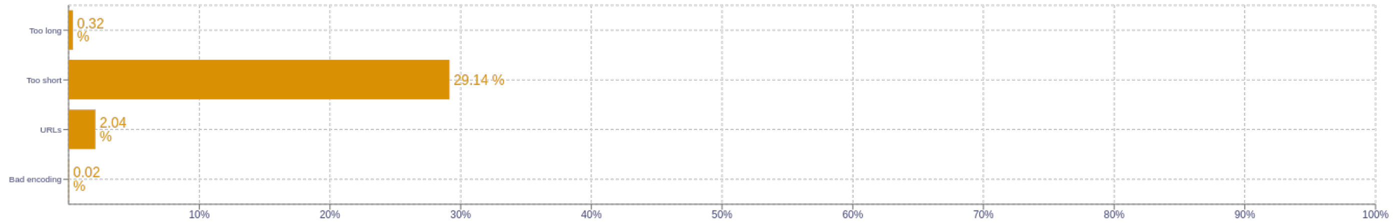


Segment length distribution by token

<= 49 tokens = **16M** segments | **58M** duplicates
> 50 tokens = **2.2M** segments | **640K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	the | 8403881 and | 5541226 of | 5116153 to | 4386027 kitabu | 3900398
2	kuweka mbadala | 1333470 hifadhi kitabu | 1323536 watch kitabu | 1323535 vitabu vyote | 1140830 of the | 1036971
3	ingizo la nyaraka | 1177419 lugha ya kiingereza | 1087104 mwaka mmoja uliopita | 515763 is licensed by | 406869 icons made by | 406869
4	made by freepik from | 406865 icons made by freepik | 406865 by freepik from www.flaticon.com | 406865 www.flaticon.com is licensed by | 406864 is licensed by cc | 406864
5	made by freepik from www.flaticon.com | 406865 icons made by freepik from | 406865 www.flaticon.com is licensed by cc | 406864 from www.flaticon.com is licensed by | 406864 freepik from www.flaticon.com is licensed | 406864

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