

General overview

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

Volumes

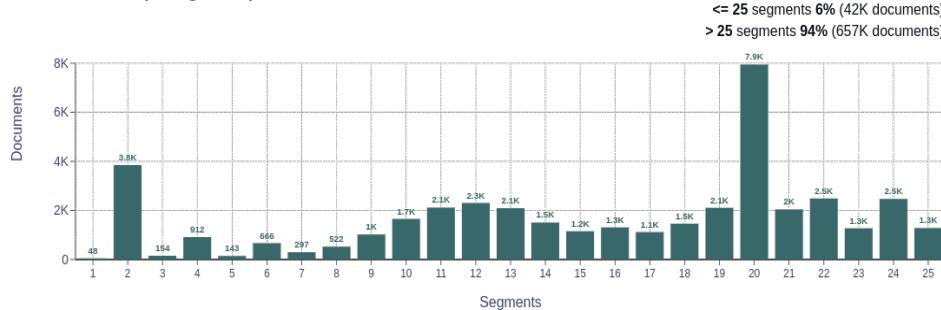
Docs	Segments	Unique segments	Tokens	Size
698,565	76,253,152	41,094 (0.05 %)	862M	4.09 GB

Type-Token Ratio

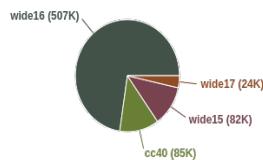
Swahili (sw)

0.01

Documents size (in segments)

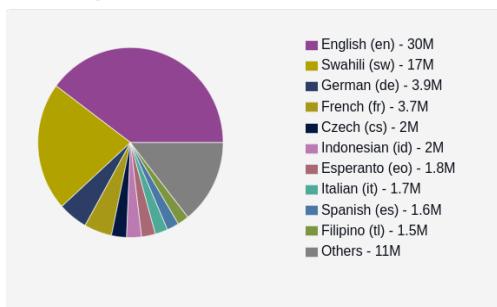


Documents by collection

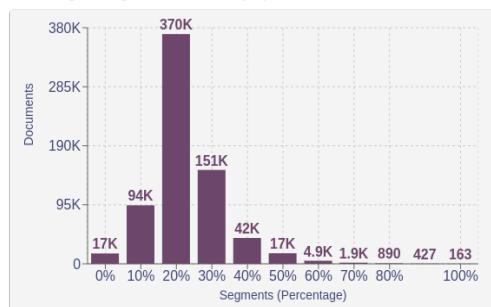


Language Distribution

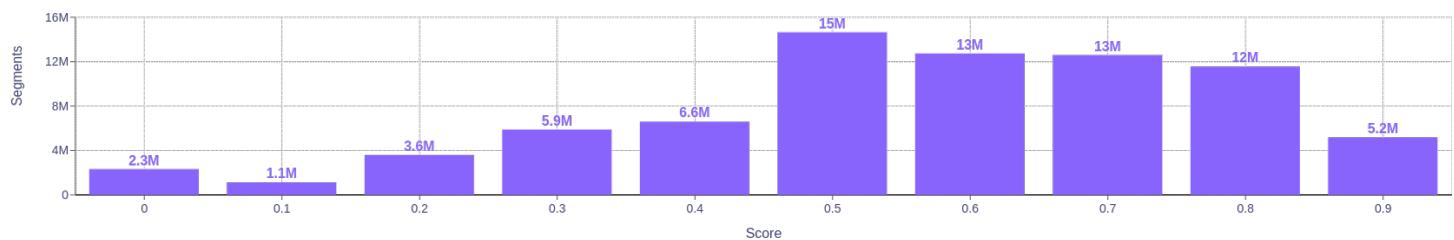
Number of segments



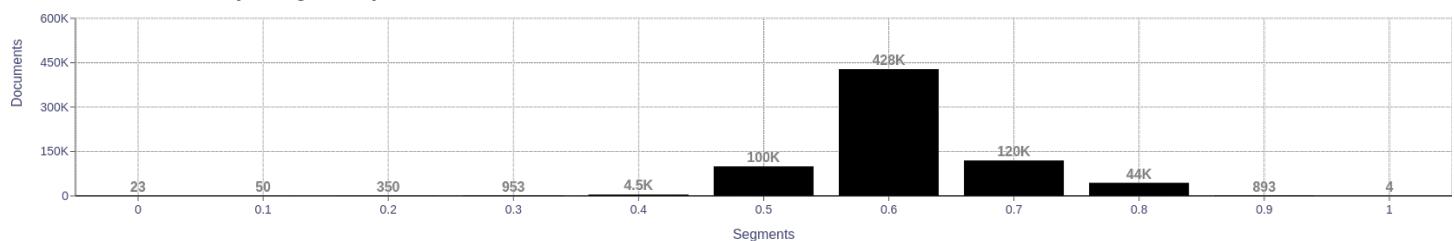
Percentage of segments in Swahili (sw) inside documents



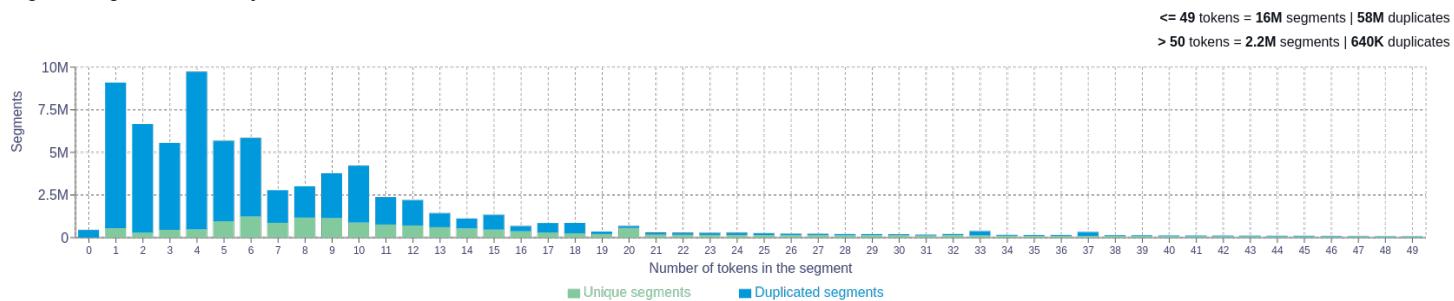
Distribution of segments by fluency score



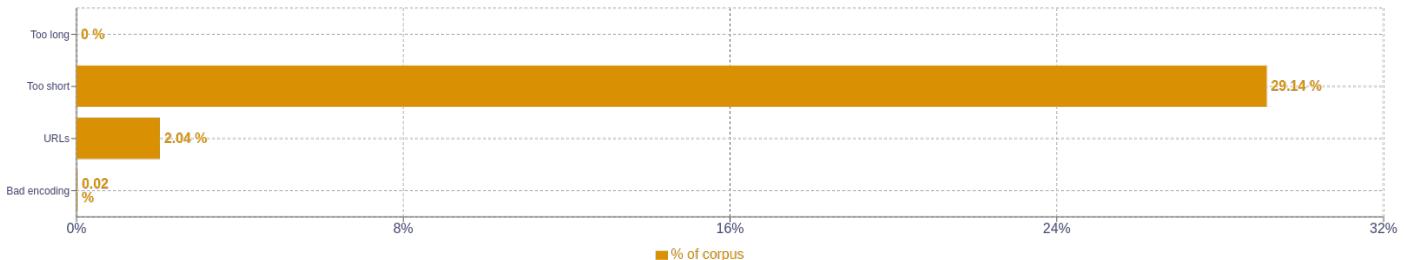
Distribution of documents by average fluency score



Segment length distribution by token



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>, , , 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>).

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.

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>