

General overview

Corpus	Analytics date	Language
swlh_Latn.jsonl.tsv	9/22/2024	Swahili (swl)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,373,860	34,308,768	14,568,570 (42.46 %)	834M	4.34 GB	4,631,243,067

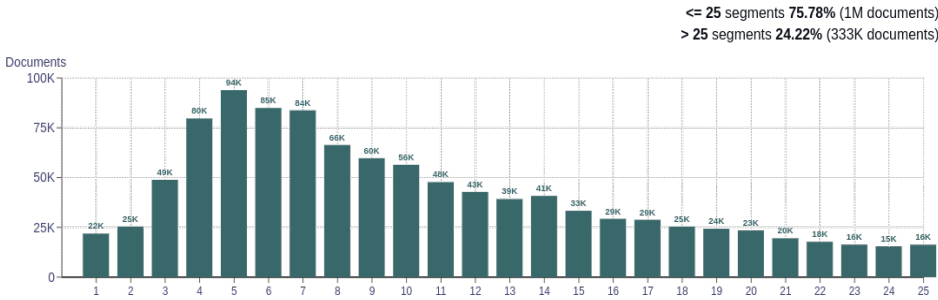
Top 10 domains

Domain	Docs	% of total
blogspot.com	190K	13.81
unmultimedia.org	65K	4.76
wikipedia.org	58K	4.19
ackyshine.com	53K	3.88
jamiiforums.com	35K	2.56
tuko.co.ke	29K	2.14
voaswahili.com	23K	1.70
mtanzania.co.tz	17K	1.21
mwanahalisonline.com	15K	1.06
dw.com	14K	1.05

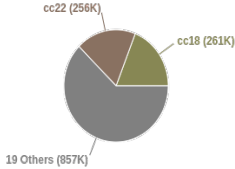
Top 10 TLDs

Domain	Docs	% of total
com	773K	56.26
org	208K	15.15
co.tz	149K	10.84
co.ke	60K	4.38
go.tz	23K	1.70
net	21K	1.56
fr	14K	1.00
no	9.3K	0.67
info	8K	0.58
com.br	6.6K	0.48

Documents size (in segments)

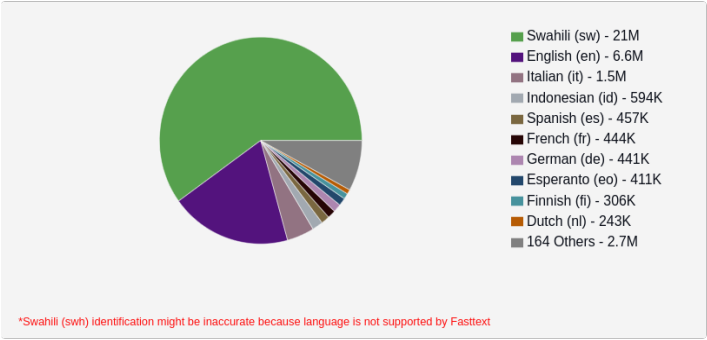


Documents by collection

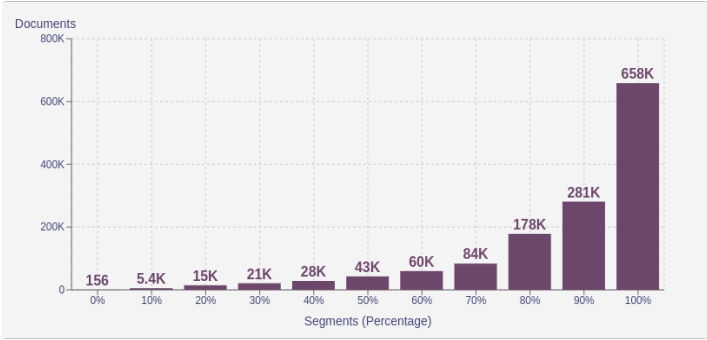


Language Distribution

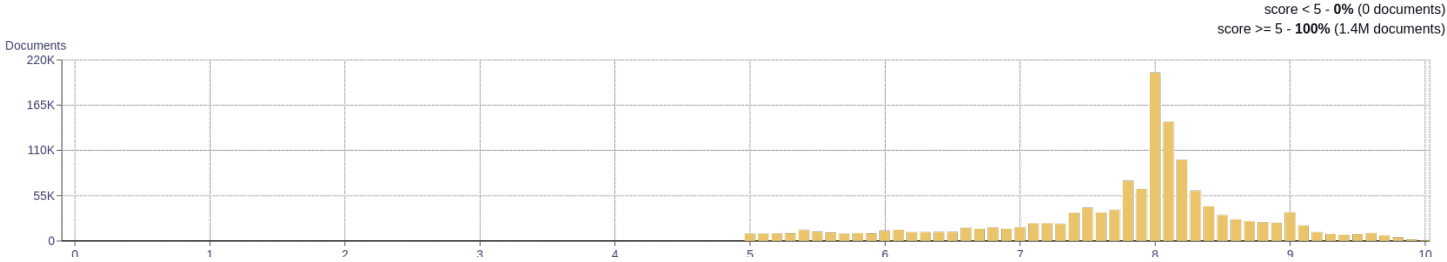
Number of segments



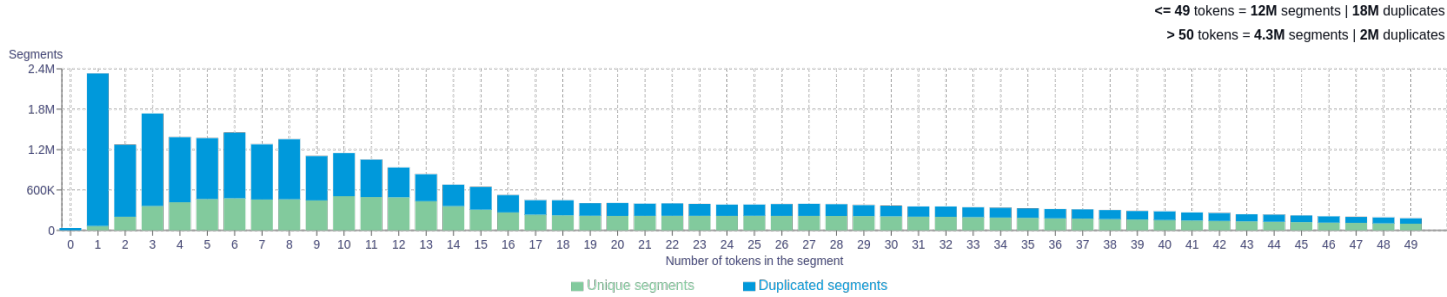
Percentage of segments in Swahili (swl) inside documents



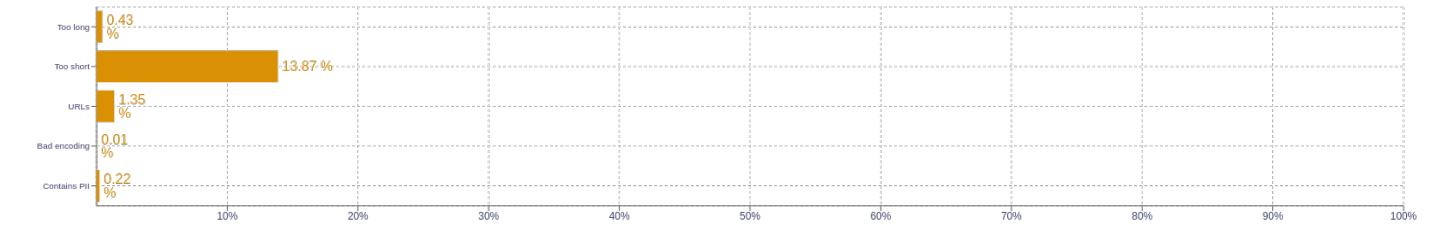
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>hii   2122035</div> <div>au   2046121</div> <div>vya   1750493</div> <div>ni   1693235</div> <div>mwaka   1437946</div>
2	<div>dar es   545049</div> <div>es salaam   535678</div> <div>mwaka huu   267434</div> <div>jijini dar   257086</div> <div>mwenyezi mungu   220984</div>
3	<div>dar es salaam   531282</div> <div>umoja wa mataifa   275814</div> <div>jijini dar es   228418</div> <div>jamhuri ya muungano   147850</div> <div>waandishi wa habari   138397</div>
4	<div>jijini dar es salaam   222328</div> <div>katibu mkuu wa umoja   41998</div> <div>ofisi ya waziri mkuu   35241</div> <div>mkoa wa dar es   28285</div> <div>dar es salaam leo   27464</div>
5	<div>jamhuri ya muungano wa tanzania   120460</div> <div>rais wa jamhuri ya muungano   81605</div> <div>mkuu wa umoja wa mataifa   40642</div> <div>akizungumza na waandishi wa habari   33455</div> <div>mwenyekiti wa baraza la mapinduzi   30186</div>

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 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>, <ul>, <ol>, etc.) replaced by newlines.

Language distribution

Language identifiied 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>