

General overview

Corpus	Date	Language
hplt-v3-swh_Latn	9/24/2025	Swahili (swh)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,935,907	43,154,287	27,265,037 (63.18 %)	1.1B	6,168,365,814	5.77 GB

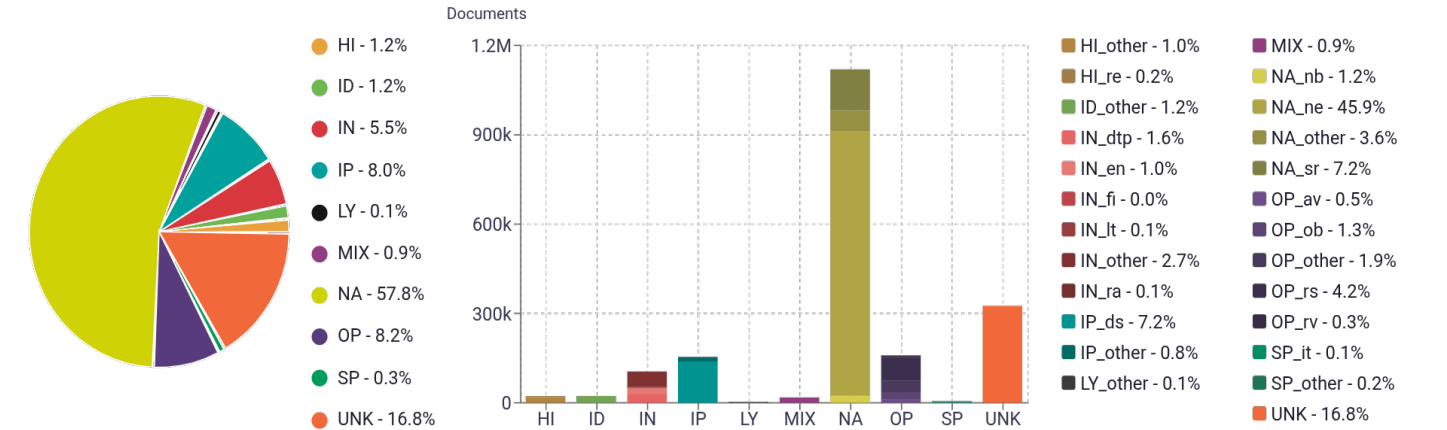
Top 10 domains

Domain	Docs	% of total
blogspot.com	148K	7.66%
airbnb.com	116K	5.97%
tuko.co.ke	56K	2.89%
nation.co.ke	31K	1.63%
mtanzania.co.tz	31K	1.61%
dw.com	30K	1.56%
millardayo.com	25K	1.27%
zanzinews.com	24K	1.22%
un.org	21K	1.08%
habarileo.co.tz	20K	1.04%

Top 10 TLDs

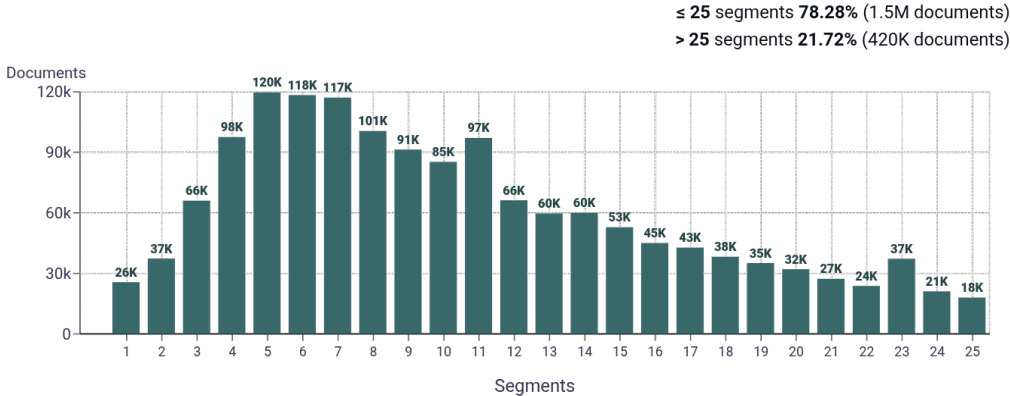
Domain	Docs	% of total
com	1.1M	54.38%
co.tz	296K	15.28%
co.ke	142K	7.33%
org	141K	7.28%
go.tz	83K	4.29%
net	32K	1.65%
fr	22K	1.12%
cn	14K	0.72%
info	11K	0.56%
va	11K	0.56%

Register labels

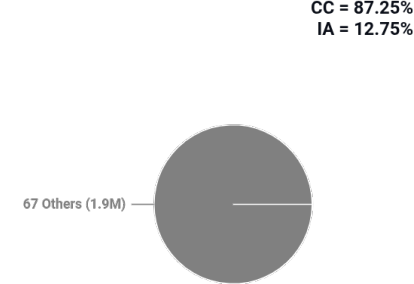


MT:17.1% | 332K Documents

Documents size (in segments)

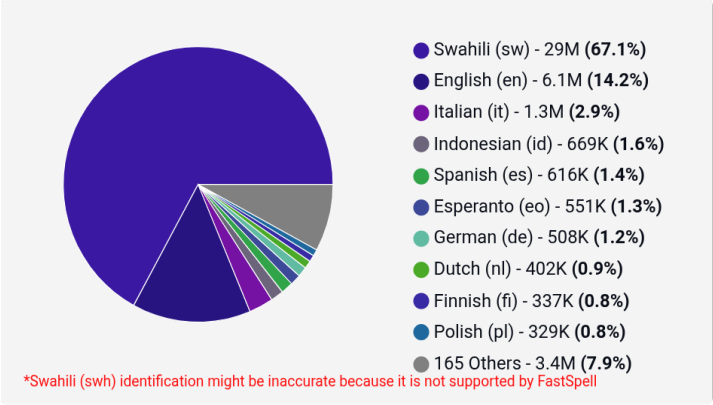


Document collections

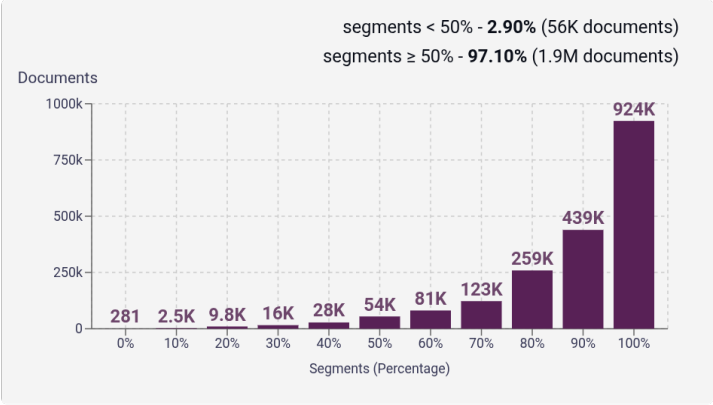


Language Distribution

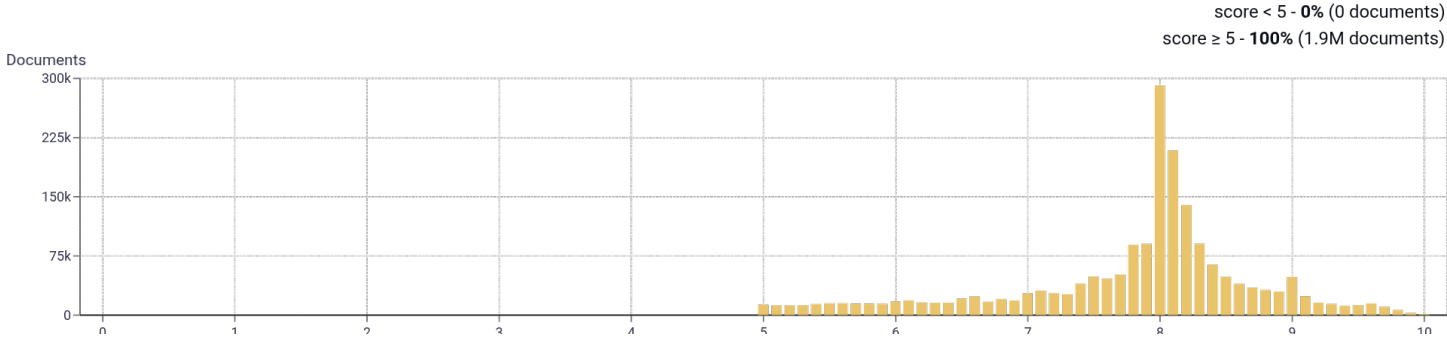
Number of segments in the Swahili (swh) corpus



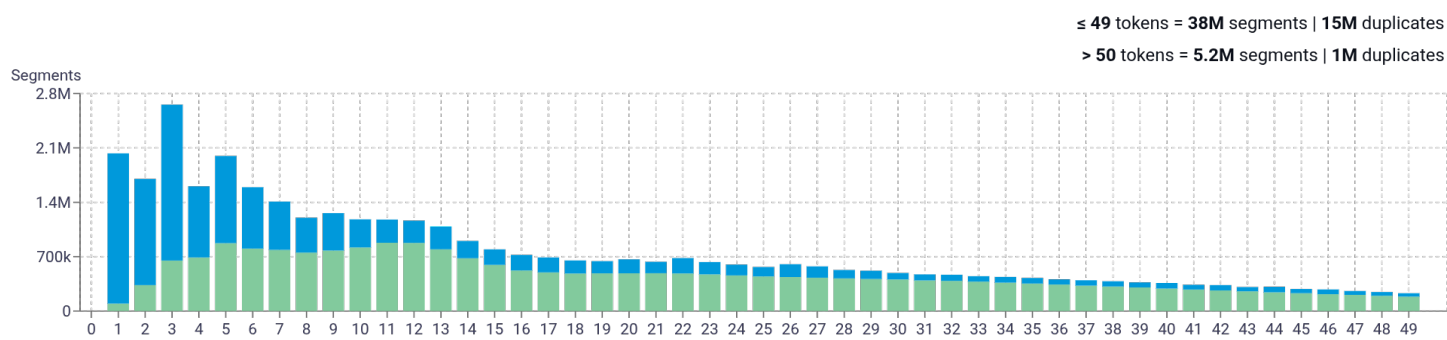
Percentage of segments in Swahili (swh) inside documents



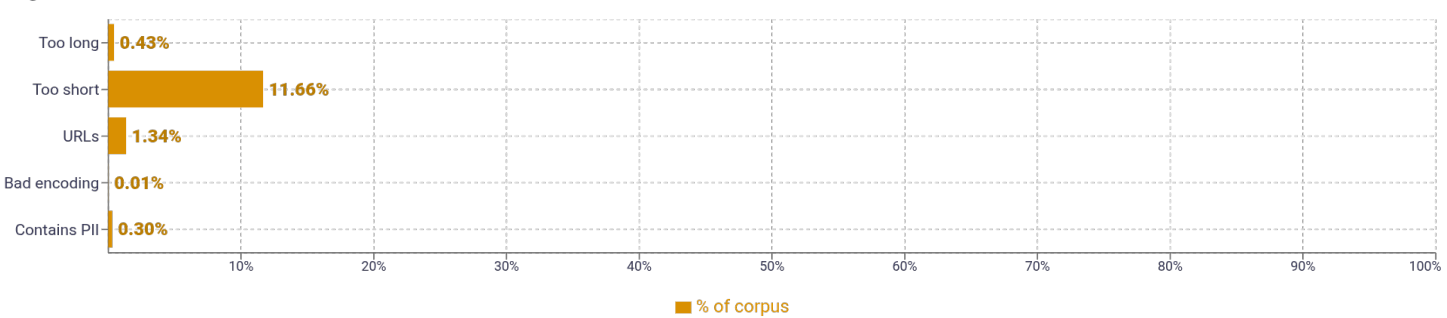
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>vya 2,917,481</div> <div>hii 2,650,121</div> <div>kazi 2,097,666</div> <div>kati 2,054,972</div> <div>ambayo 2,004,131</div>	
2	<div>dar es 580,604</div> <div>es salaam 567,271</div> <div>mwanzo huko 329,277</div> <div>mwaka huu 304,610</div> <div>kufanya kazi 275,096</div>	
3	<div>ukadiriaji wa wastani 1,006,765</div> <div>dar es salaam 561,846</div> <div>ukurasa wa mwanzo 332,022</div> <div>post a comment 242,791</div> <div>jijini dar es 218,564</div>	
4	<div>ukurasa wa mwanzo huko 329,225</div> <div>jijini dar es salaam 210,997</div> <div>nyumba ya kupangisha huko 147,945</div> <div>kukodisha wakati wa likizo 88,378</div> <div>pata na uweke nafasi 64,771</div>	
5	<div>jamhuri ya muungano wa tanzania 141,688</div> <div>rais wa jamhuri ya muungano 84,749</div> <div>pata nafasi ambayo ni sahihi 63,166</div> <div>nafasi ambayo ni sahihi kwako 63,166</div> <div>mwenyekiti wa baraza la mapinduzi 50,189</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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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>

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	How-to or instructions	HI	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	Informational persuasion	IP	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	Opinion	OP
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	Informational description	IN	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				