HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
swh_Latn.jsonl.tsv	9/22/2024	Swahili (swh)

Volumes

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

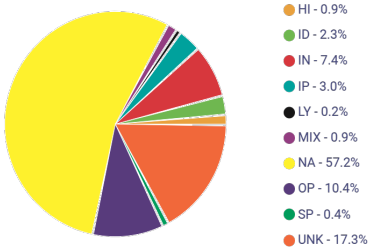
Top 10 domains

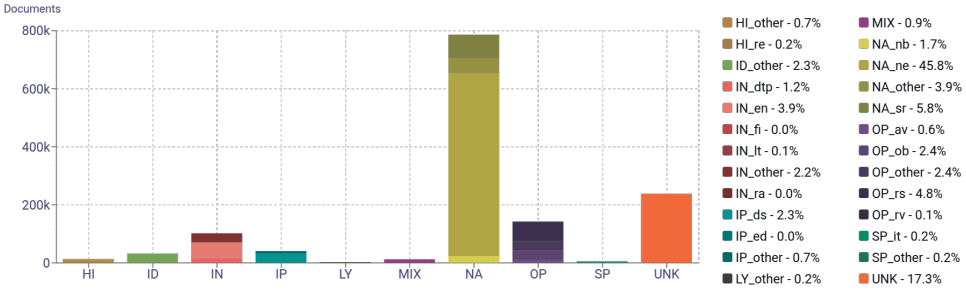
Domain	Docs	% of total
blogspot.com	190K	13.81%
unmultimedia.org	65K	4.76%
wikipedia.org	58K	4.19%
ackyshine.com	53K	3.88%
jamiiforums.com	35K	2.56%
tuko.co.ke	29K	2.14%
voaswahili.com	23K	1.70%
mtanzania.co.tz	17K	1.21%
mwanahalisionli...	15K	1.06%
dw.com	14K	1.05%

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%

Register labels

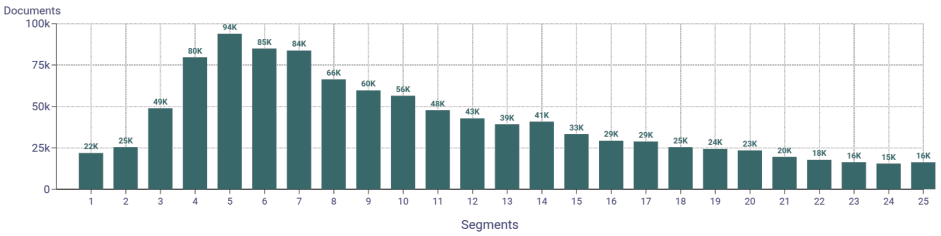




🤖 **MT**:11.3% | 155K Documents

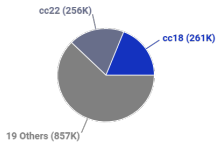
Documents size (in segments)

<= 25 segments **75.78%** (1M documents)
> 25 segments **24.22%** (333K documents)



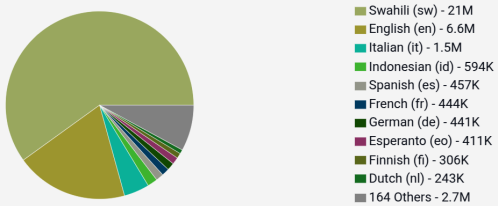
Documents by collection

CC = 57.40%
IA = 42.60%



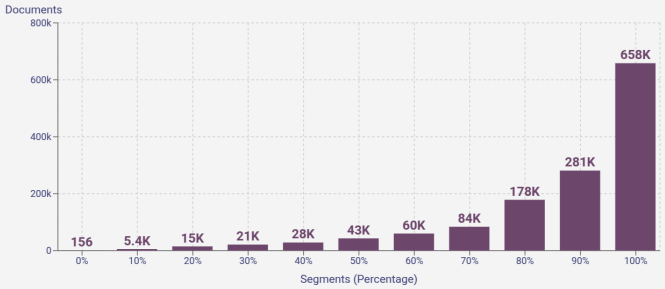
Language Distribution

Number of segments in the Swahili (swh) corpus



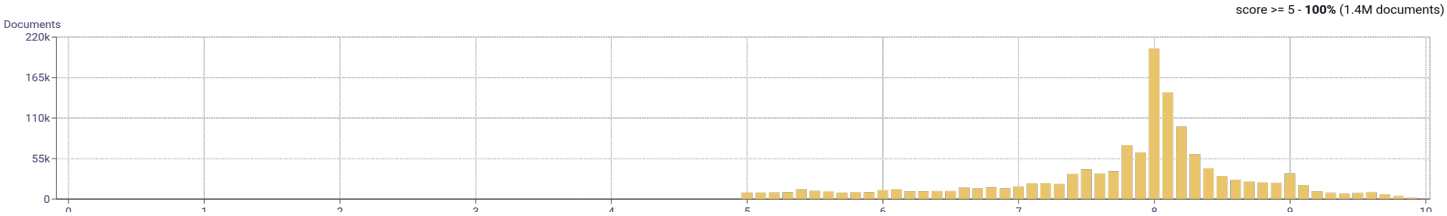
*Swahili (swh) identification might be inaccurate because it is not supported by FastSpell

Percentage of segments in Swahili (swh) inside documents



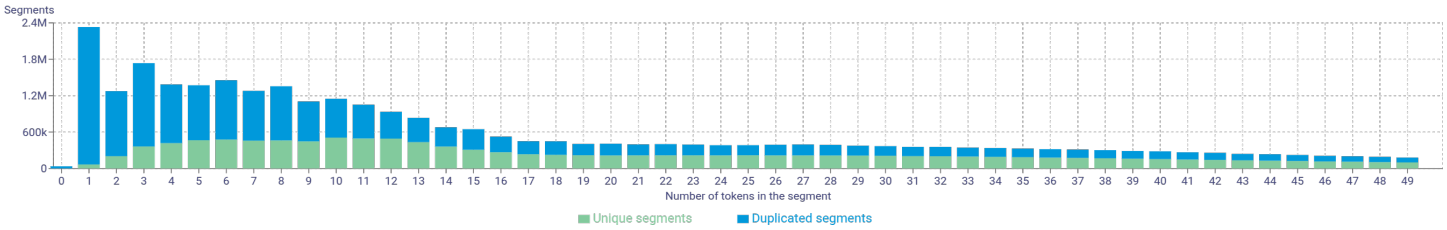
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (1.4M documents)

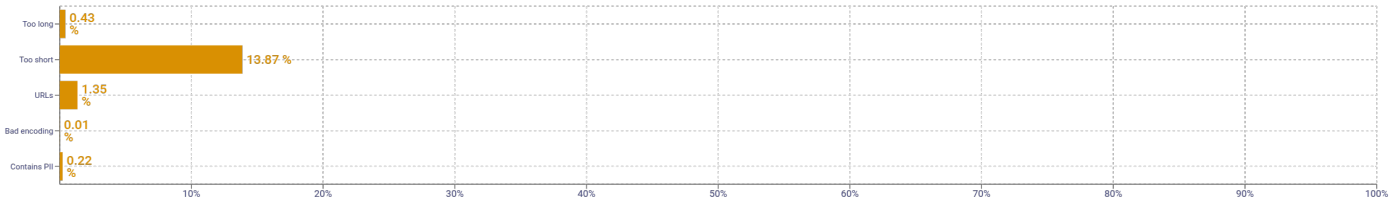


Segment length distribution by token

≤ 49 tokens = **12M** segments | **18M** duplicates
> 50 tokens = **4.3M** segments | **2M** duplicates



Segment noise distribution



Frequent n-grams

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

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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av