HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
kin_Latn.jsonl.tsv	9/23/2024	Kinyarwanda (rw)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
92,699	1,916,933	1,163,101 (60.68 %)	65M	354.87 MB	365,284,983

Top 10 domains

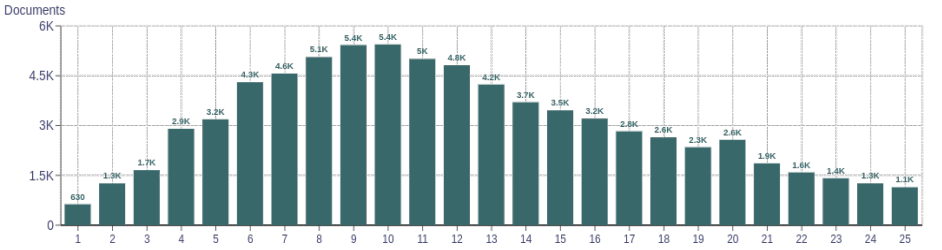
Domain	Docs	% of total
igihe.com	11K	11.38
kigalitoday.com	4.5K	4.88
agakiza.org	3.8K	4.09
jw.org	2.7K	2.87
yezu-akuzwe.org	2.6K	2.79
newsofrwanda.com	2.4K	2.64
inyarwanda.com	2.3K	2.48
agasaro.com	2.3K	2.44
imirasire.com	2.1K	2.32
umuryango.rw	2K	2.18

Top 10 TLDs

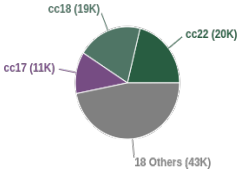
Domain	Docs	% of total
com	54K	58.54
org	17K	18.60
rw	13K	13.82
fr	2.3K	2.43
co.rw	1.5K	1.64
gov.rw	1.4K	1.48
net	1K	1.11
info	453	0.49
be	232	0.25
ca	215	0.23

Documents size (in segments)

<= 25 segments **82.52%** (76K documents)
> 25 segments **17.48%** (16K documents)

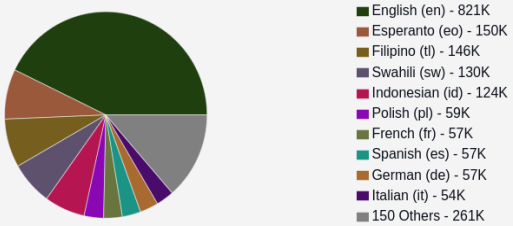


Documents by collection



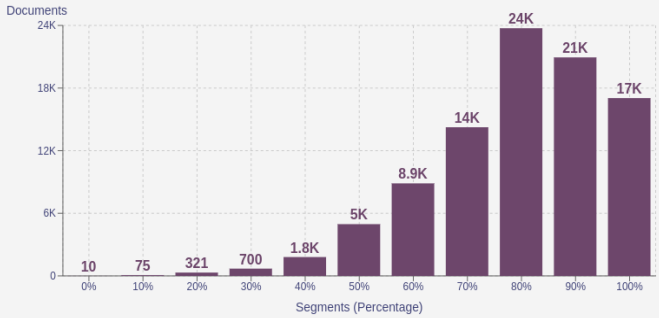
Language Distribution

Number of segments



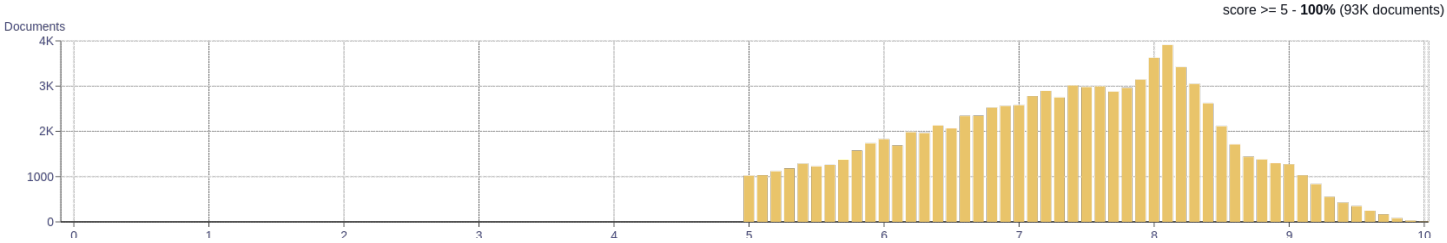
*Kinyarwanda (rw) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Kinyarwanda (rw) inside documents



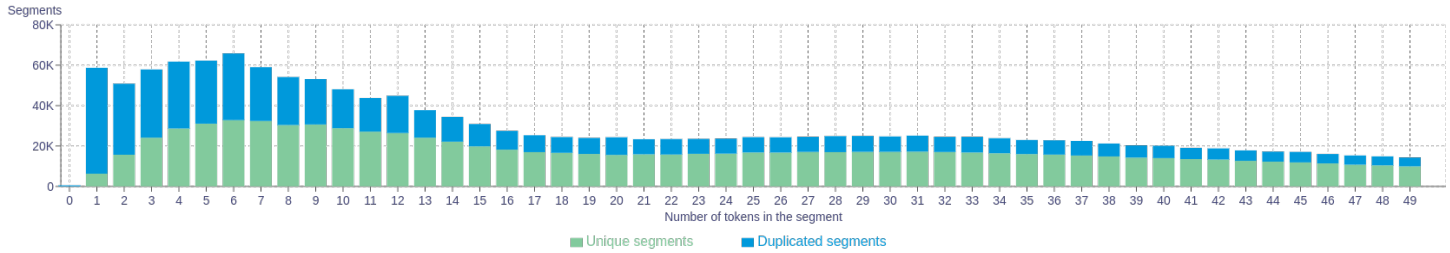
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (93K documents)

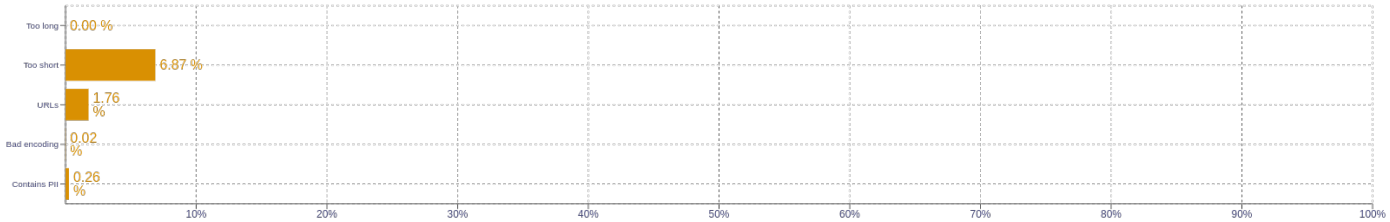


Segment length distribution by token

<= 49 tokens = **894K** segments | **617K** duplicates
> 50 tokens = **406K** segments | **137K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	n | 1103217 · y | 455189 · kandi | 364482 · imana | 319325 · w | 249241
2	u rwanda | 56569 · ndetse n | 41497 · nyuma y | 26384 · cyangwa se | 25683 · cyane cyane | 21242
3	kanda hano umusubize | 17580 · hirya no hino | 8877 · rimwe na rimwe | 8137 · jenoside yakorewe abatutsi | 6784 · mujyi wa kigali | 5388
4	bite bite bite bite | 3498 · imana imuhe amahoro n | 2147 · zunze ubumwe za amerika | 1721 · allah amuhe amahoro n | 1355 · ushinzwe imibereho myiza y | 1324
5	bite bite bite bite bite | 3493 · hirya no hino ku isi | 1735 · leta zunze ubumwe za amerika | 1700 · hirya no hino mu gihugu | 1581 · alayhi wa aalih wa sallam | 1249

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