HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
run_Latn.jsonl.tsv	9/20/2024	Rundi (rn)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
137,296	1,751,899	1,093,620 (62.42 %)	56M	306.26 MB	314,875,644

Top 10 domains

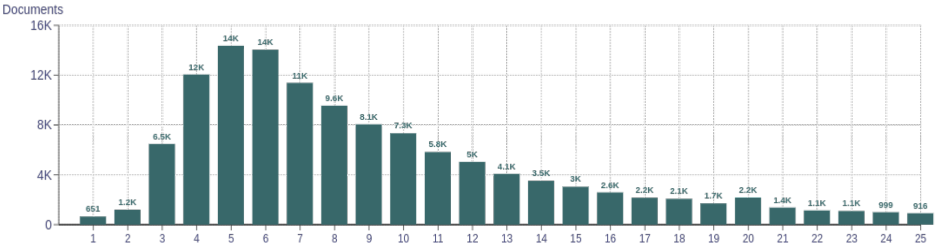
Domain	Docs	% of total
radiyoyacuvoa.com	16K	11.77
igihe.bi	15K	11.29
igihe.com	8.1K	5.93
indundi.com	7.6K	5.55
ruhagoyacu.com	5.7K	4.19
umuryango.rw	4.2K	3.08
jw.org	4K	2.93
kigalitoday.com	3.3K	2.38
bbc.com	2.8K	2.06
veritasinfo.fr	2.7K	1.99

Top 10 TLDs

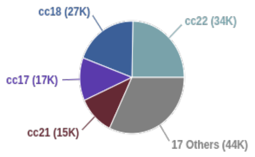
Domain	Docs	% of total
com	78K	57.08
rw	24K	17.12
bi	16K	12.00
org	7.5K	5.45
fr	3.9K	2.83
net	2.6K	1.87
co.rw	1.9K	1.35
is	647	0.47
info	399	0.29
gov.rw	227	0.17

Documents size (in segments)

<= 25 segments **89.48%** (123K documents)
> 25 segments **10.52%** (14K documents)

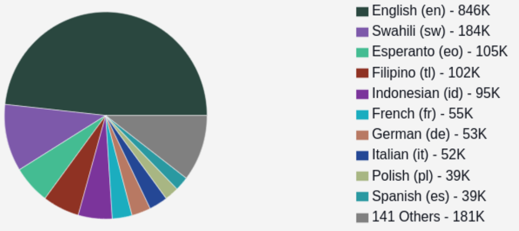


Documents by collection



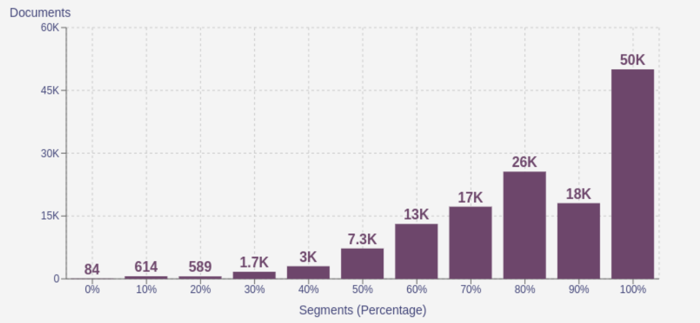
Language Distribution

Number of segments



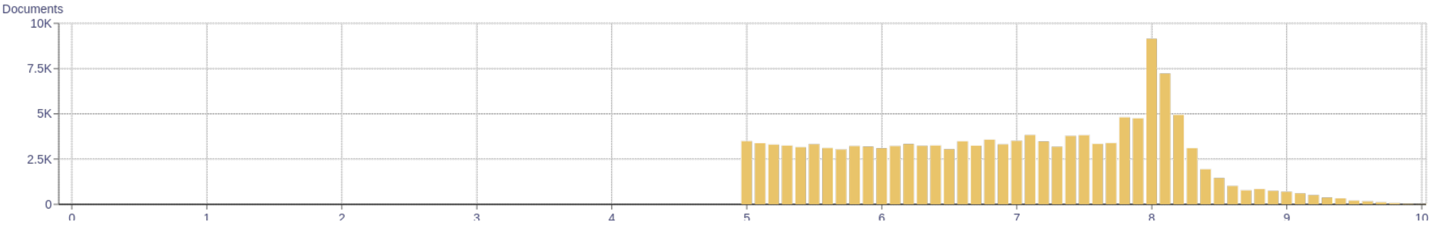
*Rundi (rn) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Rundi (rn) inside documents



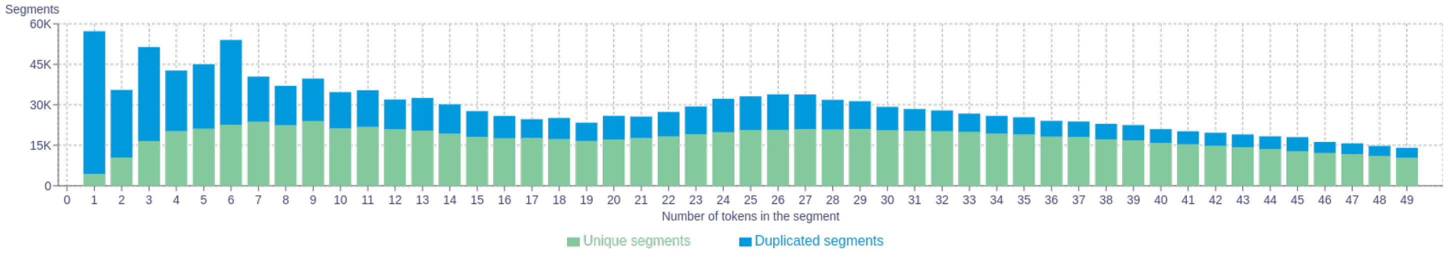
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (137K documents)

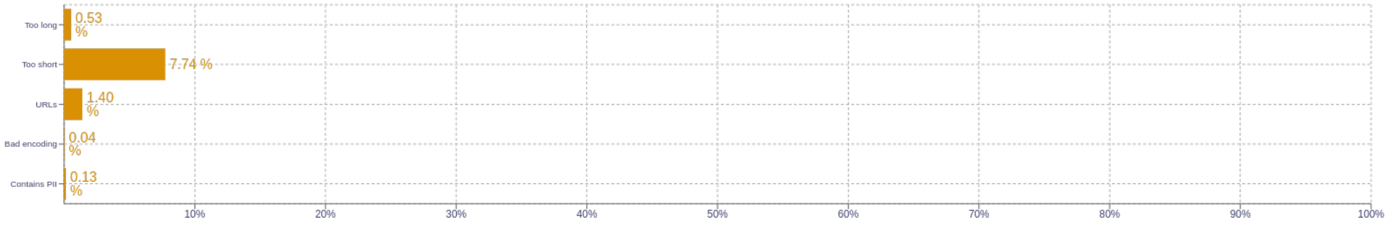


Segment length distribution by token

<= 49 tokens = **874K** segments | **559K** duplicates
> 50 tokens = **319K** segments | **100K** duplicates



Segment noise distribution



Frequent n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>n | 882735
y | 552799
w | 375977
rwanda | 214931
kandi | 180657</td></tr>
<tr><td>2</td><td>u rwanda | 108819
umukuru w | 44525
nyuma y | 40371
rayon sports | 36415
hamwe n | 28686</td></tr>
<tr><td>3</td><td>leta zunze ubumwe | 14506
ubumwe za amerika | 12351
jenoside yakorewe abatutsi | 8405
zunze ubumwe z | 7172
ikipe ya rayon | 6775</td></tr>
<tr><td>4</td><td>zunze ubumwe za amerika | 12110
leta zunze ubumwe z | 5733
ikipe ya rayon sports | 5096
iharanira demokarasi ya congo | 4941
ikipe ya apr fc | 4066</td></tr>
<tr><td>5</td><td>leta zunze ubumwe za amerika | 7970
repubulika iharanira demokarasi ya congo | 4409
reta zunze ubumwe za amerika | 4091
urutonde rwa batsinze ibitego byinshi | 2505
urutonde rurambuye rwa ba rutahizamu | 2505</td></tr>
</table>

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