HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
yor_Latn.jsonl.tsv	9/21/2024	Yoruba (yo)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
66,132	1,468,729	989,734 (67.39 %)	50M	241.64 MB	216,421,805

Top 10 domains

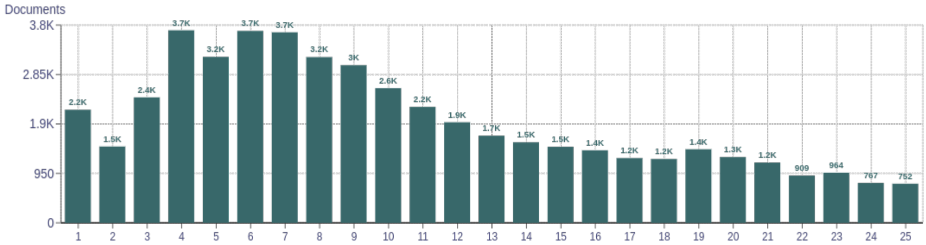
Domain	Docs	% of total
alaroye.org	4.7K	7.15
vessoft.com	3K	4.49
awikonko.com.ng	2.3K	3.43
wikipedia.org	2.2K	3.33
ilorin.info	1.8K	2.67
jw.org	1.6K	2.36
creativosonline.org	1.5K	2.22
androidsis.com	1.3K	2.04
martech.zone	1.2K	1.86
bible.is	1.1K	1.65

Top 10 TLDs

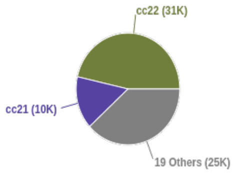
Domain	Docs	% of total
com	38K	56.80
org	13K	19.37
com.ng	3.1K	4.66
info	2.2K	3.35
net	1.8K	2.77
zone	1.2K	1.86
is	1.1K	1.68
top	856	1.29
es	474	0.72
co.uk	392	0.59

Documents size (in segments)

<= 25 segments **74.14%** (49K documents)
> 25 segments **25.86%** (17K documents)

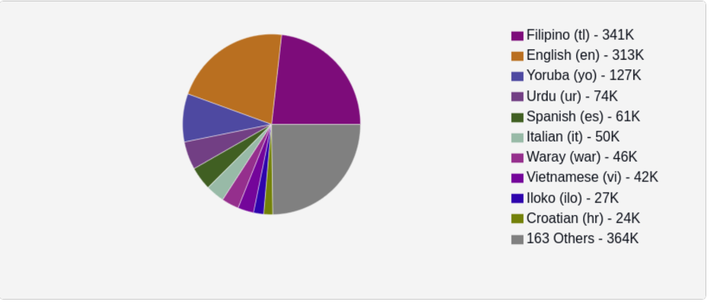


Documents by collection

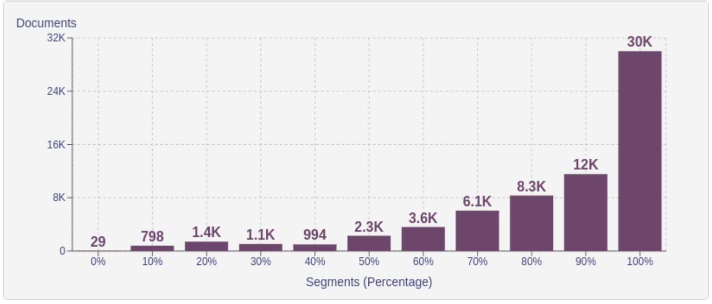


Language Distribution

Number of segments

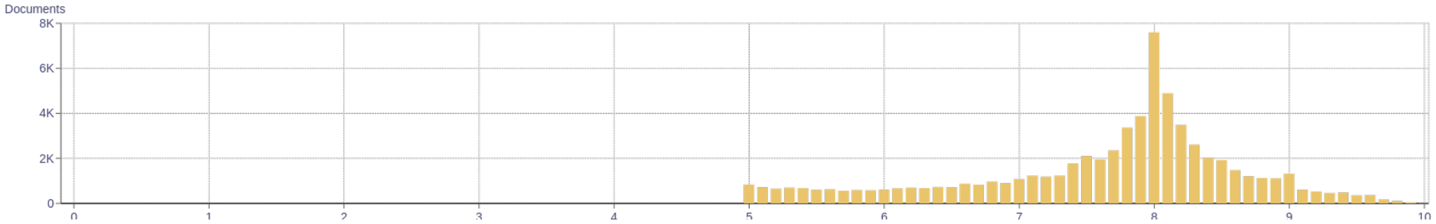


Percentage of segments in Yoruba (yo) inside documents



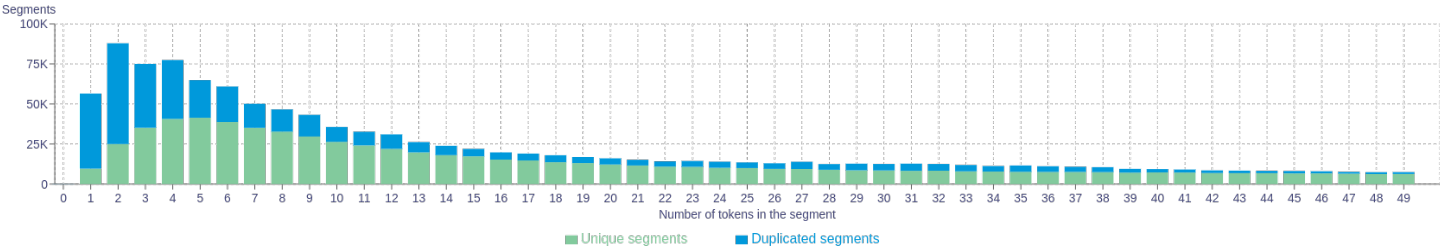
Distribution of documents by document score

score <= 5 - **99.97%** (66K documents)
score > 5 - **0.03%** (19 documents)

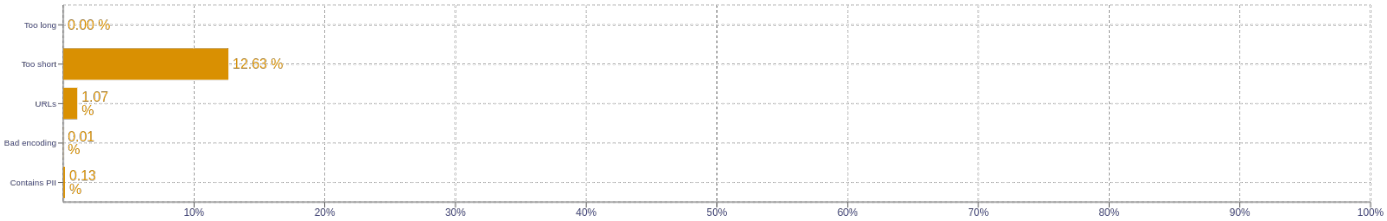


Segment length distribution by token

<= 49 tokens = **723K** segments | **424K** duplicates
> 50 tokens = **322K** segments | **55K** duplicates



Segment noise distribution



Frequent n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>awọn | 1862034
ati | 790440
lati | 642410
si | 508363
fun | 455459</td></tr>
<tr><td>2</td><td>ati awọn | 144358
fun awọn | 87241
ninu awọn | 74375
ohun elo | 74186
pẹlu awọn | 66569</td></tr>
<tr><td>3</td><td>ki o si | 38853
awọn ohun elo | 33190
diẹ ninu awọn | 25802
ọkan ninu awọn | 21047
awọn ẹya ara | 12412</td></tr>
<tr><td>4</td><td>awọn ẹya ara ẹrọ | 10347
faye gba o lati | 10052
awọn software faye gba | 6389
bii o ṣe le | 5705
wo diẹ sii software | 5457</td></tr>
<tr><td>5</td><td>software faye gba o lati | 6781
òßí òßí òßí òßí òßí | 4304
awọn ti o dara ju | 3738
fun fun fun fun fun | 3381
akọkọ awọn ẹya ara ẹrọ | 2452</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