HPLT Analytics report

General overview

Corpus	Analytics date	Language
ibo_Latn.jsonl.tsv	12/6/2024	Igbo (ig)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
56,288	1,411,417	971,706 (68.85 %)	47M	229.43 MB	203,796,797

Top 10 domains

Domain	Docs	% of total
jw.org	6K	10.71
vessoft.com	2.8K	5.06
martech.zone	1.9K	3.41
wikipedia.org	1K	1.78
bbc.com	744	1.32
kaoditaa.com	712	1.26
igboradio.com	626	1.11
actualidadviajes.com	623	1.11
androidsis.com	606	1.08
eturbonews.com	585	1.04

Top 10 TLDs

Domain	Docs	% of total
com	36K	64.56
org	11K	18.77
zone	1.9K	3.41
net	1.4K	2.50
ru	1.1K	2.03
tv	505	0.90
com.ng	498	0.88
dk	359	0.64
fr	322	0.57
co.uk	278	0.49

Documents size (in segments)

<= 25 segments **70.31%** (40K documents)
> 25 segments **29.69%** (17K documents)
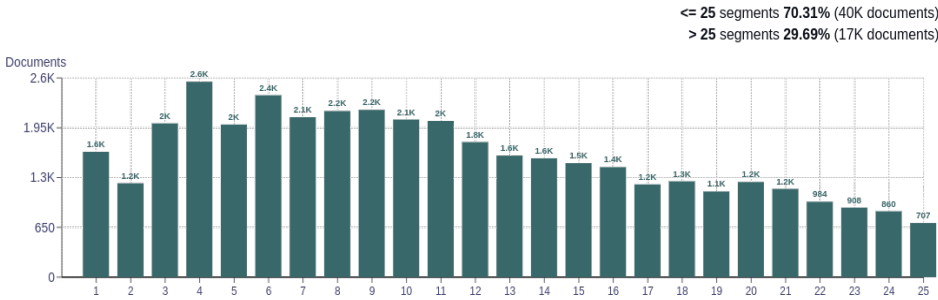
Documents by collection

cc22 (24K)
cc21 (9K)
19 Others (23K)
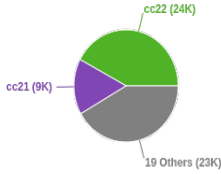
Language Distribution

Number of segments

- English (en) - 439K
- Vietnamese (vi) - 400K
- Croatian (hr) - 106K
- Filipino (tl) - 97K
- Italian (it) - 41K
- French (fr) - 32K
- German (de) - 25K
- Swahili (sw) - 23K
- Spanish (es) - 16K
- Haitian Creole (ht) - 12K
- 161 Others - 221K

*Igbo (ig) identification might be inaccurate because language is not supported by Fasttext
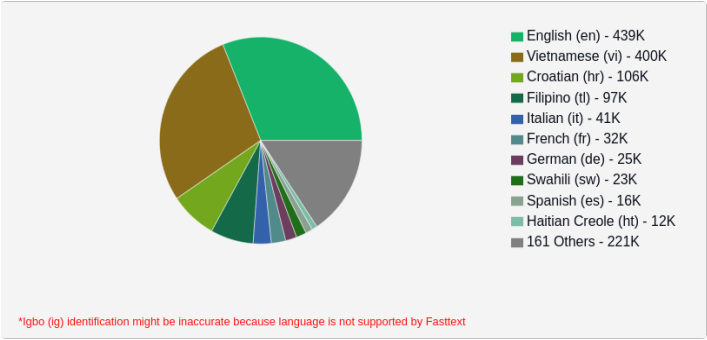
Percentage of segments in Igbo (ig) inside documents
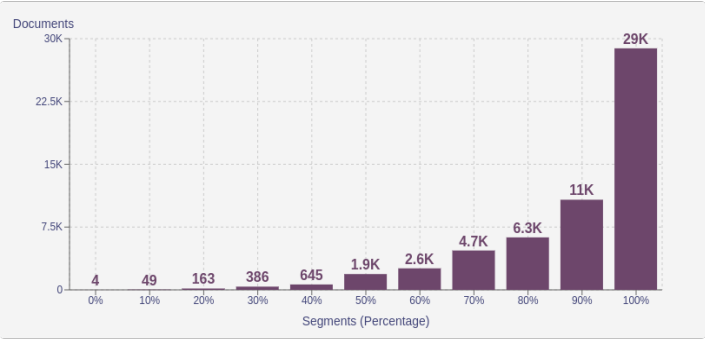
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (56K documents)

Segment length distribution by token

<= 49 tokens = **710K** segments | **387K** duplicates
> 50 tokens = **314K** segments | **52K** duplicates

- Unique segments
- Duplicated segments

Segment noise distribution

Too long	0.55 %
Too short	12.78 %
URLs	1.19 %
Bad encoding	0.00 %
Contains PII	0.11 %
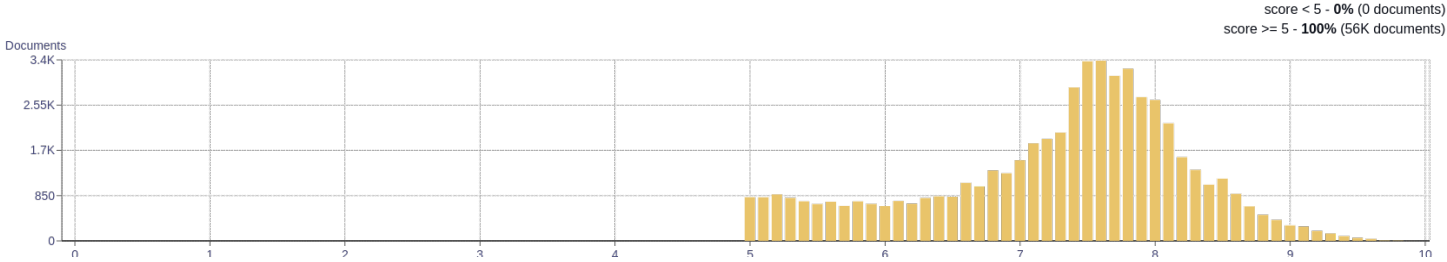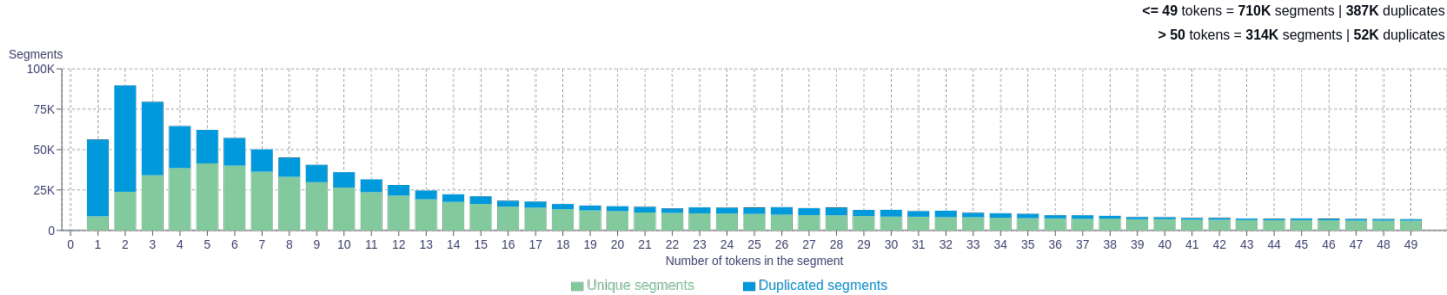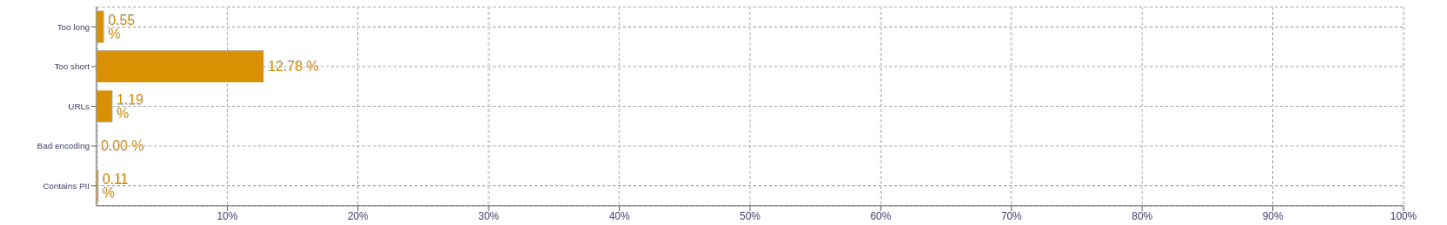
Frequent n-grams

Size	n-grams
1	na- | 1266269 n' | 692489 ndị | 631747 ya | 493217 ike | 384300
2	nwere ike | 161924 na- eme | 76694 ndị na- | 76423 ndị ọzọ | 75476 iche iche | 72694
3	na- arụ ọrụ | 33096 onye ọ bụla | 18490 enye gị ohere | 15128 anyị nwere ike | 13358 ole na ole | 12394
4	bụrụ na ị na- | 10227 akụkọ ihe mere eme | 6750 na- aga n' ihu | 6743 nwa nwa nwa nwa | 5855 na- enye gị ohere | 5458
5	nwa nwa nwa nwa nwa | 5680 the software- enye gị ohere | 5009 na- arụ ọrụ nke ọma | 1940 na- eme ka ọ bụrụ | 1566 ndị na- eso ụzọ ya | 1553

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