HPLT Analytics report

General overview

Corpus	Analytics date	Language
fuv_Latn.jsonl.tsv	10/31/2024	Nigerian Fulfulde (fuv)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
7,760	133,977	70,149 (52.36 %)	6.3M	29.93 MB	29,772,757

Top 10 domains

Domain	Docs	% of total
pulaar.org	2.9K	37.74
bible.is	941	12.13
fuutamedia.com	512	6.60
wikipedia.org	414	5.34
rfi.fr	296	3.81
binndipulaar.com	271	3.49
jw.org	219	2.82
lewlewal.com	214	2.76
misiide.net	183	2.36
wol-children.net	170	2.19

Top 10 TLDs

Domain	Docs	% of total
org	3.9K	49.92
com	1.6K	21.04
is	941	12.13
net	596	7.68
fr	313	4.03
ir	159	2.05
gov.ng	124	1.60
in	38	0.49
info	29	0.37
ru	8	0.10

Documents size (in segments)

<= 25 segments **82.65%** (6.4K documents)
> 25 segments **17.35%** (1.3K documents)
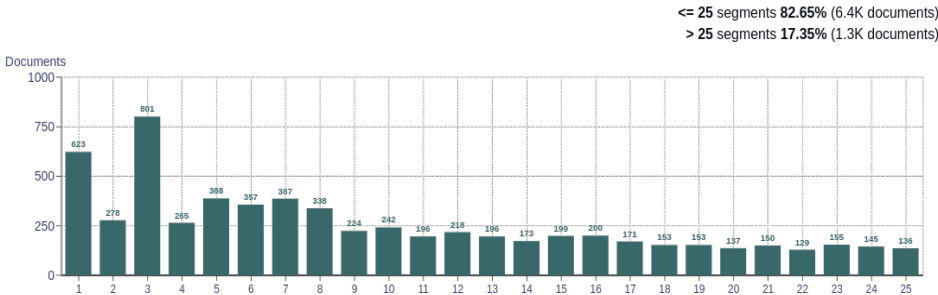
Documents by collection

cc22 (2.1K), cc18 (1.3K), cc21 (1K), wide12 (777), 17 Others (2.5K)
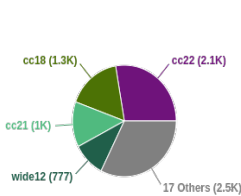
Language Distribution

Number of segments

- English (en) - 41K
- Finnish (fi) - 27K
- Welsh (cy) - 8.3K
- French (fr) - 7.9K
- Latin (la) - 5.6K
- Estonian (et) - 5.4K
- Dutch (nl) - 4.5K
- Italian (it) - 4.1K
- German (de) - 2.9K
- Spanish (es) - 2.8K
- 135 Others - 24K

*Nigerian Fulfulde (fuv) identification might be inaccurate because language is not supported by Fasttext
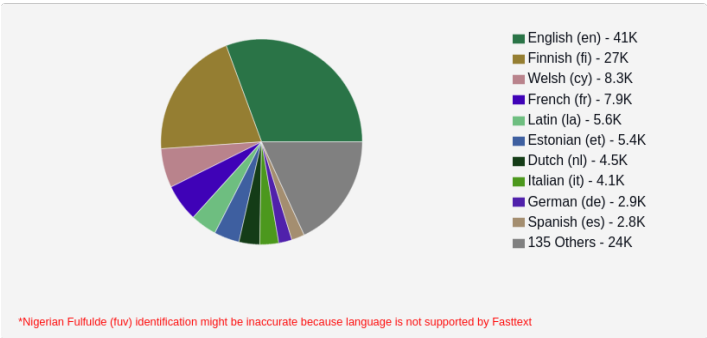
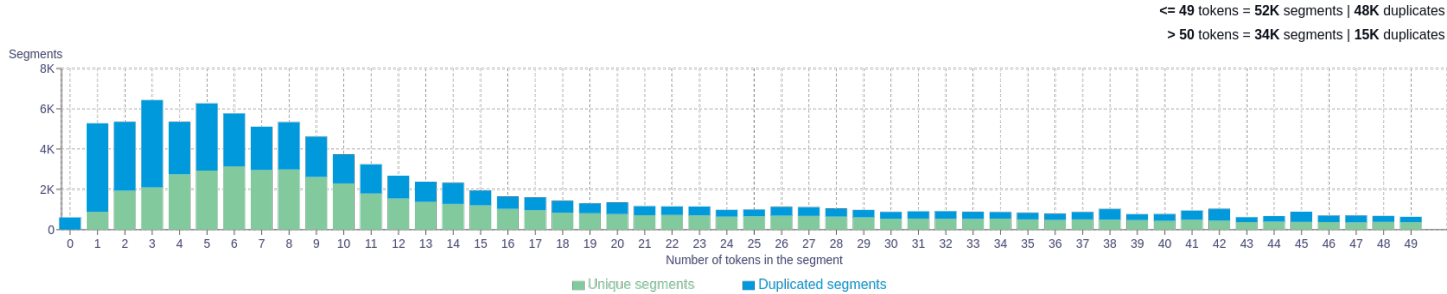### Percentage of segments in Nigerian Fulfulde (fuv) inside documents
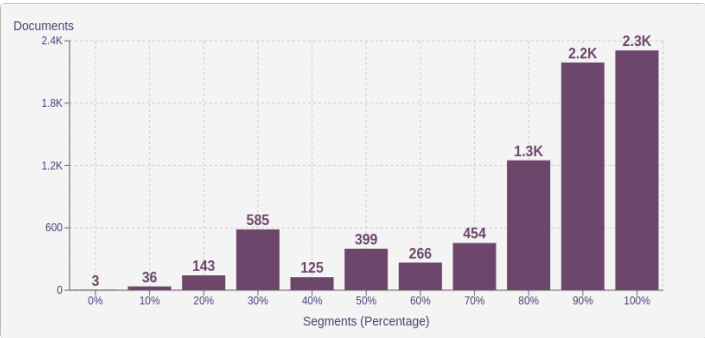
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (7.8K documents)

Segment length distribution by token

<= **49** tokens = **52K** segments | **48K** duplicates
> **50** tokens = **34K** segments | **15K** duplicates

Number of tokens in the segment

- Unique segments
- Duplicated segments

Segment noise distribution

- Too long: 0.00 %
- Too short: 5.10 %
- URLs: 1.04 %
- Bad encoding: 0.00 %
- Contains PII: 0.08 %
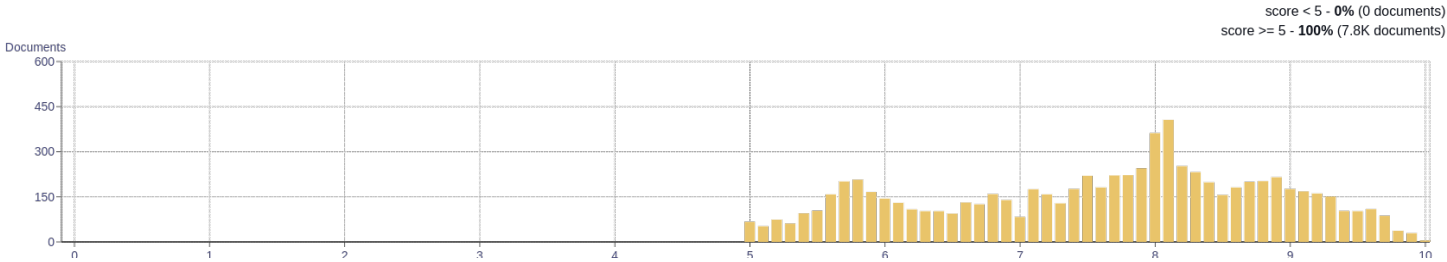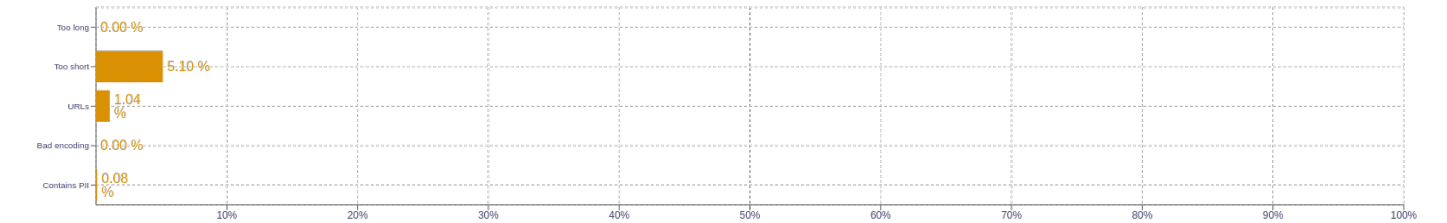
About HPLT Analytics

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