

General overview

Corpus	Analytics date	Language
fuv_Latn.jsonl.tsv	12/3/2024	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

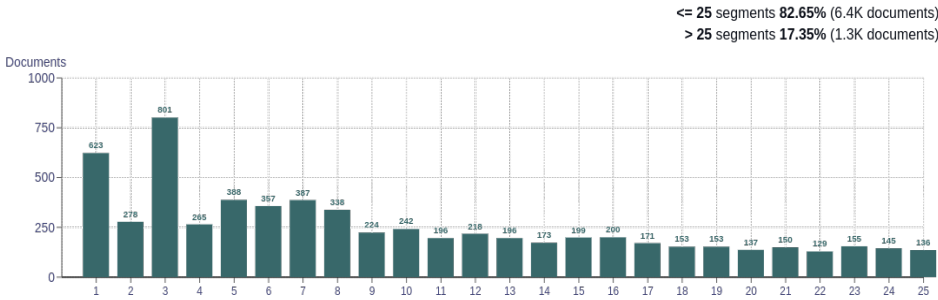
Top 10 domains

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
msiide.net	183	2.36
wol-children.net	170	2.19

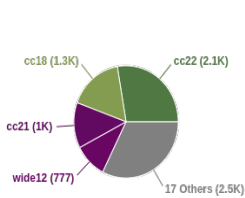
Top 10 TLDs

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)

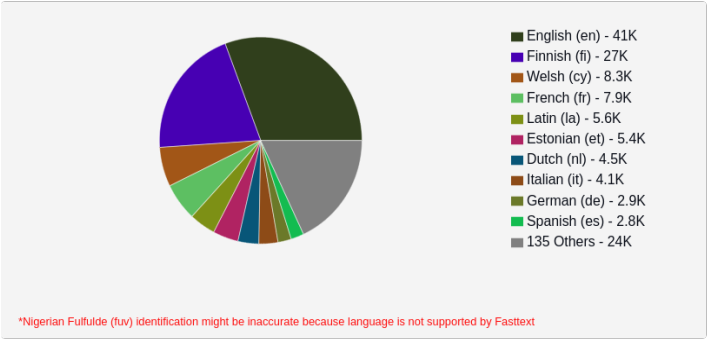


Documents by collection

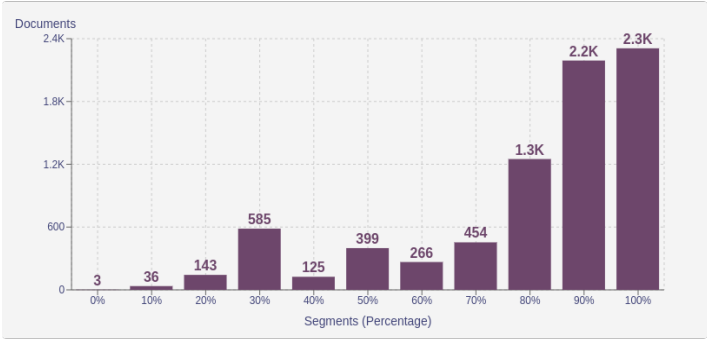


Language Distribution

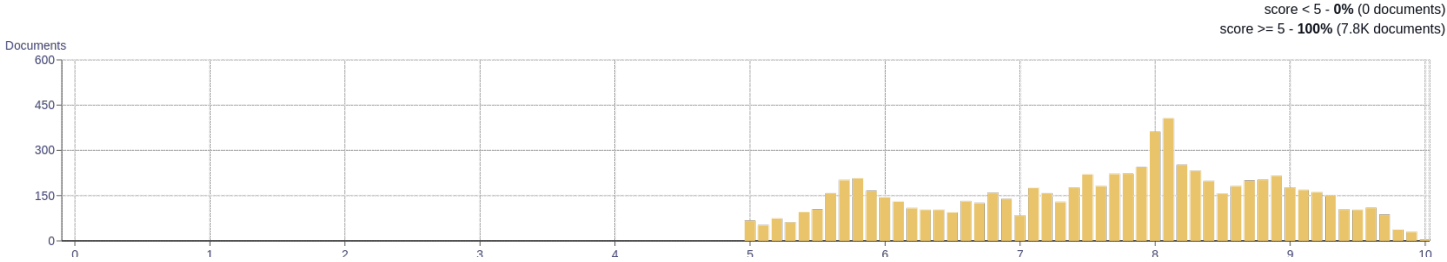
Number of segments



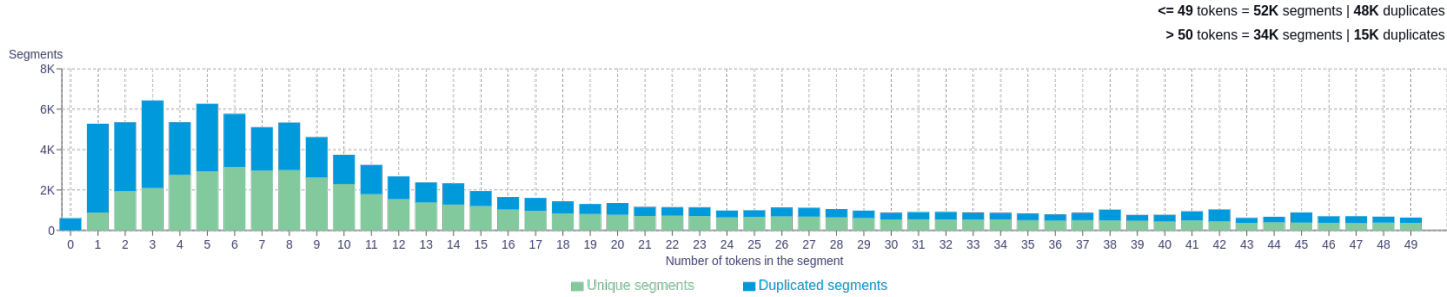
Percentage of segments in Nigerian Fulfulde (fuv) inside documents



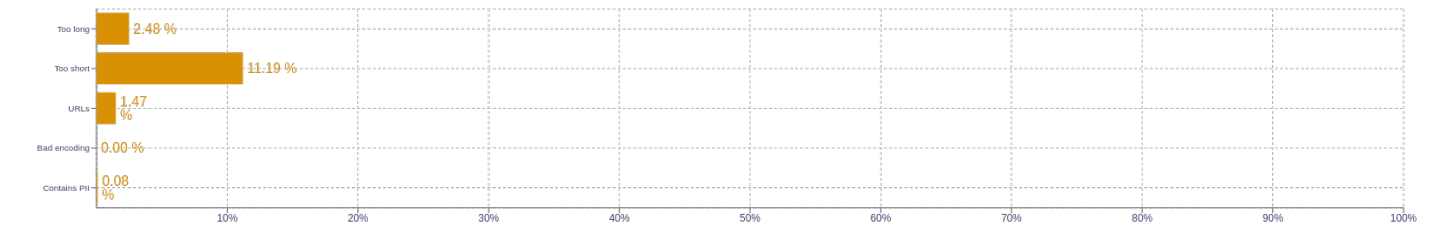
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>ki   37412</div> <div>di   33858</div> <div>ina   32566</div> <div>mum   31492</div> <div>nder   26945</div>
2	<div>so tawii   4027</div> <div>leydi ndii   3485</div> <div>mbii di   2625</div> <div>di wuri   2380</div> <div>nyem di   2223</div>
3	<div>hay so tawii   917</div> <div>bookara aamadu bah   774</div> <div>alaa e sago   677</div> <div>fedde bantaare pulaar   593</div> <div>sat mmo nee   522</div>
4	<div>wuri sat mmo nee   361</div> <div>kl kl kl kl   350</div> <div>teeco taar raa ki   327</div> <div>taar teeco taar raa   314</div> <div>taad taar teeco taar   299</div>
5	<div>kl kl kl kl kl   345</div> <div>taar teeco taar raa ki   301</div> <div>taad taar teeco taar raa   299</div> <div>fedde bantaare pulaar e muritani   290</div> <div>di ba mo a nyem   213</div>

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 of 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>