HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
hau_Latn.jsonl.tsv	9/20/2024	Hausa (ha)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
315,870	5,688,420	3,787,469 (66.58 %)	180M	820.35 MB	848,139,069

Top 10 domains

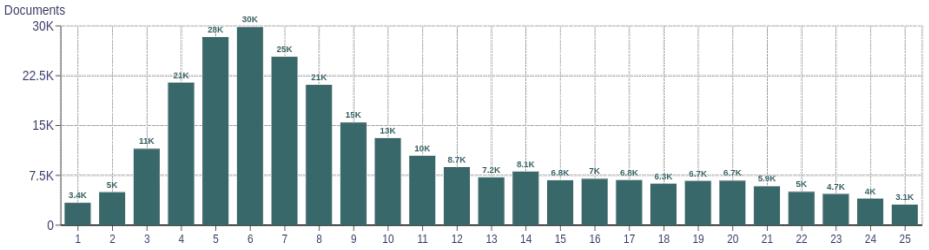
Domain	Docs	% of total
voahausa.com	48K	15.16
legit.ng	33K	10.45
leadership.ng	23K	7.27
premiumtimesng.com	18K	5.74
rfi.fr	8.5K	2.68
bbc.com	8.4K	2.65
cri.cn	6.4K	2.01
dw.com	5K	1.60
isyaku.com	4.5K	1.43
wondershare.com	4.5K	1.43

Top 10 TLDs

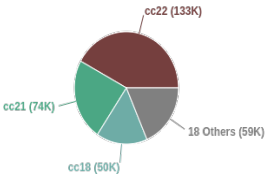
Domain	Docs	% of total
com	190K	60.15
ng	61K	19.34
com.ng	19K	6.14
org	11K	3.46
fr	9.3K	2.94
cn	6.5K	2.07
net	4K	1.26
ir	3.6K	1.14
zone	1.8K	0.56
co.uk	1.3K	0.40

Documents size (in segments)

<= 25 segments **86.05%** (272K documents)
> 25 segments **13.95%** (44K documents)

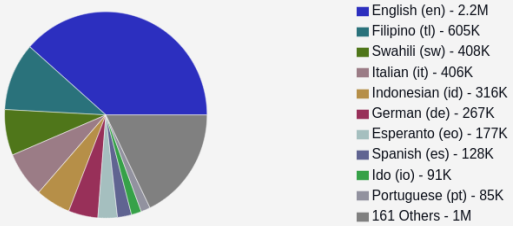


Documents by collection



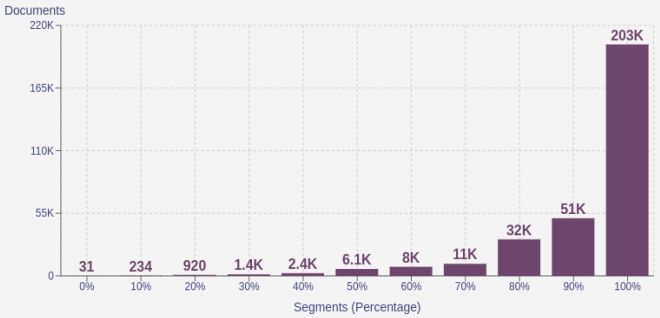
Language Distribution

Number of segments



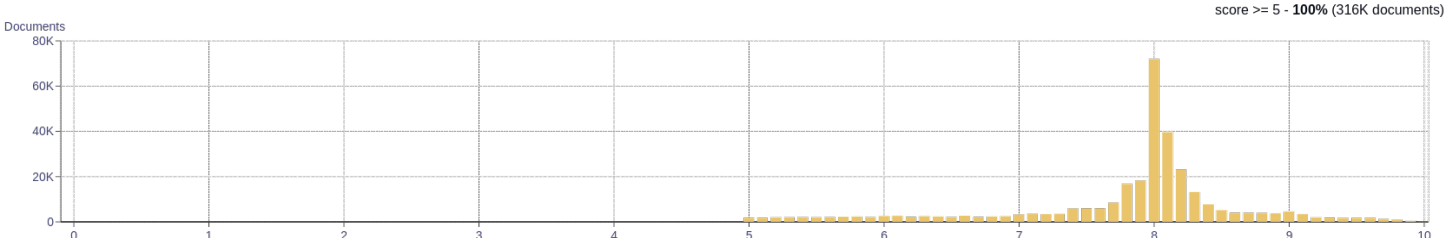
*Hausa (ha) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Hausa (ha) inside documents



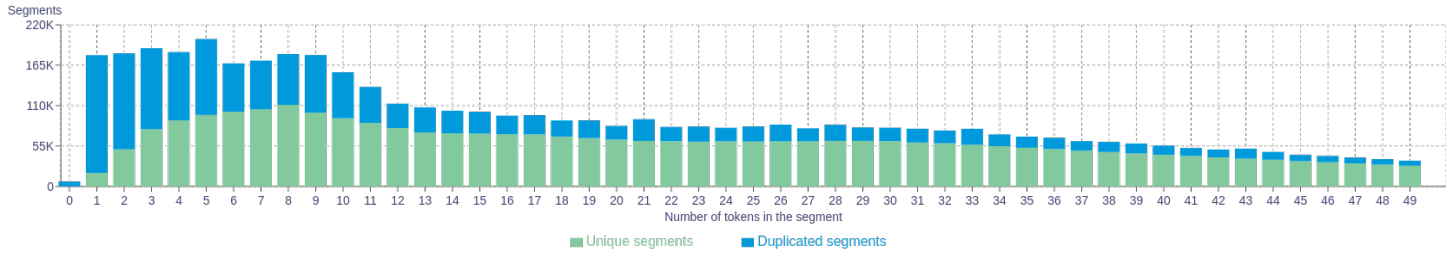
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (316K documents)

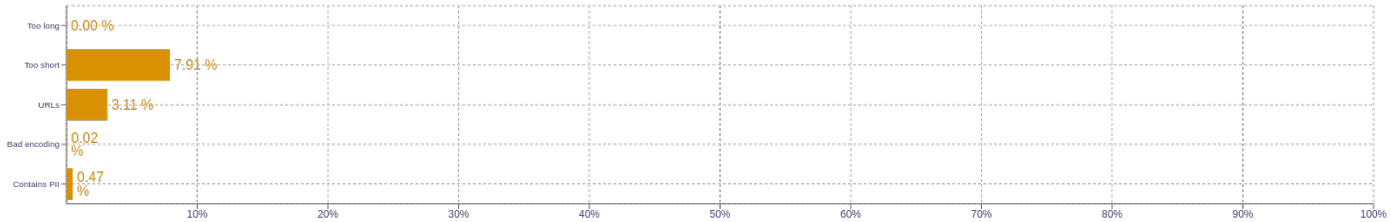


Segment length distribution by token

<= 49 tokens = **3M** segments | **1.7M** duplicates
> 50 tokens = **987K** segments | **212K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	tare | 443713 · haka | 427231 · kasar | 414901 · allah | 355204 · jihar | 339309
2	ci gaba | 139703 · shugaban kasa | 79477 · boko haram | 41826 · kayan aiki | 40359 · gwamnan jihar | 39630
3	majalisar dinkin duniya | 17795 · kasa da kasa | 15999 · dandalin sada zumunta | 13811 · kasa muhammadu buhari | 13176 · shugaban kasa muhammadu | 13173
4	shugaban kasa muhammadu buhari | 12988 · wayar ku ta hannu | 12395 · shawara ko bukatar bamu | 11826 · shafukanmu na dandalin sada | 11675 · latsa wannan domin samun | 9175
5	shawara ko bukatar bamu labari | 11825 · shafukanmu na dandalin sada zumunta | 11675 · latsa wannan domin samun sabuwar | 8055 · sabuwar manhajar labarai ta legit | 7418 · ng a shafinka na facebook | 5274

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