HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
kbp_Latn.jsonl.tsv	12/5/2024	Kabiyè (kbp)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
7,075	46,792	26,333 (56.28 %)	5.3M	24.47 MB	20,864,671

Top 10 domains

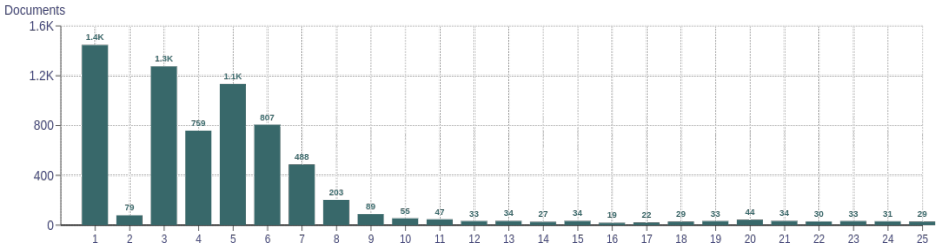
Domain	Docs	% of total
wikipedia.org	3.8K	53.87
bible.is	2.4K	34.57
jw.org	488	6.90
ebible.org	89	1.26
bibles.org	84	1.19
breakeveryyoke.com	33	0.47
wikiplanet.click	28	0.40
bywiki.com	12	0.17
revue-gugu.org	9	0.13
unicode.org	6	0.08

Top 10 TLDs

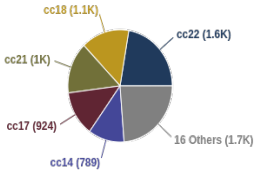
Domain	Docs	% of total
org	4.5K	63.70
is	2.4K	34.57
com	71	1.00
click	28	0.40
net	13	0.18
cf	6	0.08
ca	2	0.03
fr	1	0.01
nl	1	0.01

Documents size (in segments)

<= 25 segments **96.33%** (6.8K documents)
> 25 segments **3.67%** (260 documents)

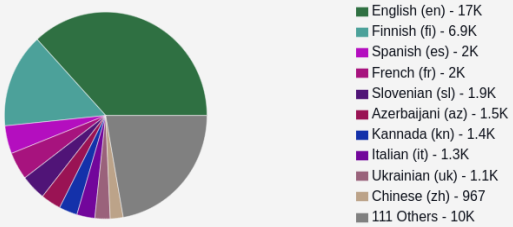


Documents by collection



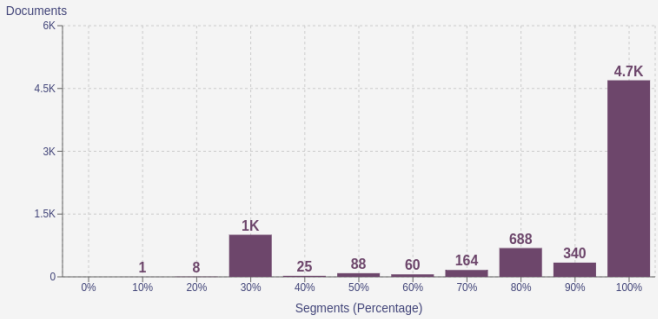
Language Distribution

Number of segments



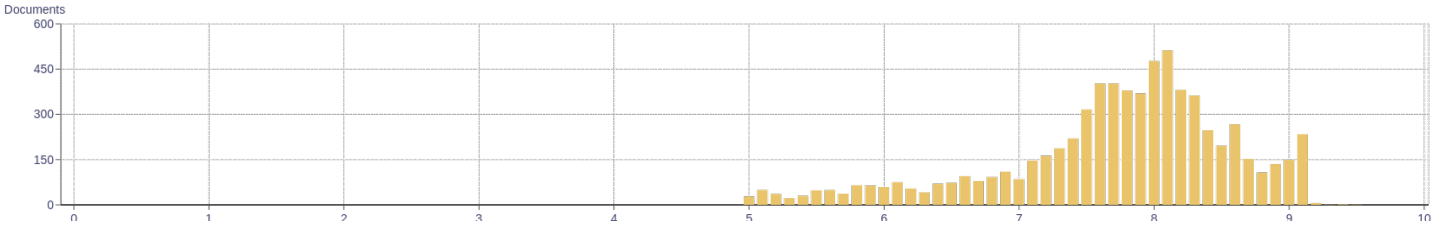
*Kabiyè (kbp) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Kabiyè (kbp) inside documents



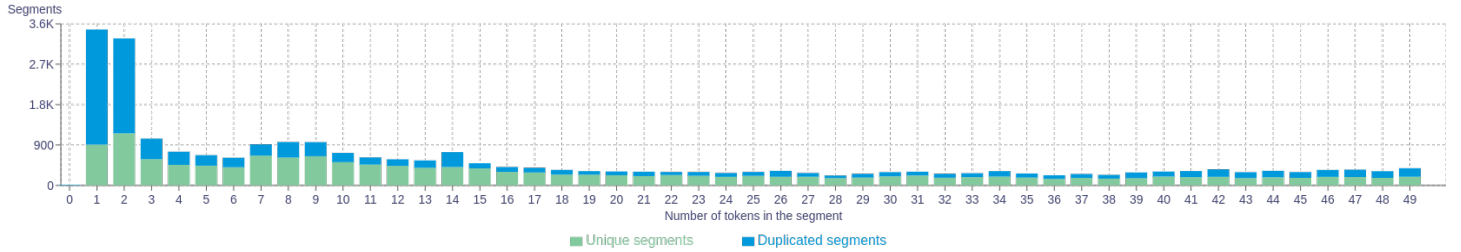
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (7.1K documents)

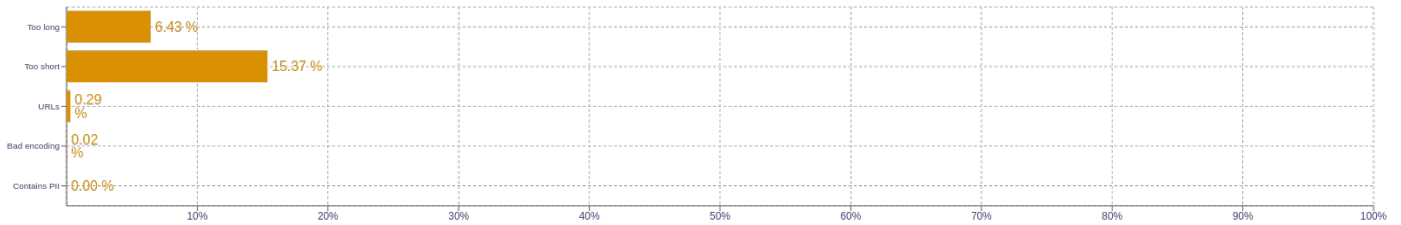


Segment length distribution by token

<= 49 tokens = **15K** segments | **12K** duplicates
> 50 tokens = **20K** segments | **8.9K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ɪ | 91417 · taa | 78909 · ba | 64121 · sɪ | 53301 · ma | 53182
2	hʊ aa | 5303 · dɪ ba | 4766 · dɪ ʊ | 3980 · ɛjaɖɛ taa | 3948 · te tɪn | 3921
3	nala hʊ aa | 1545 · pə taɣa pʊlʊ | 1492 · ɛjaɖɛ ɖɪnɛ ɖɪ | 1242 · hʊ bulonʲ aa | 1062 · ʊ a baa | 956
4	tɪya ba a baa | 687 · pə taɣa pʊlʊ tɔɔ | 687 · tɪya ʊ a baa | 608 · yee yee yee yee | 570 · rɛ yesu basɪ tɪya | 432
5	basɪ tɪya ba a baa | 639 · basɪ tɪya ʊ a baa | 604 · yee yee yee yee yee | 568 · kalʊ na ba ti tɪn | 327 · avr mai juin juil août | 311

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