HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
bam_Latn.jsonl.tsv	10/3/2024	Bambara (bm)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
5,721	91,722	42,851 (46.72 %)	4.9M	21.25 MB	20,651,317

Top 10 domains

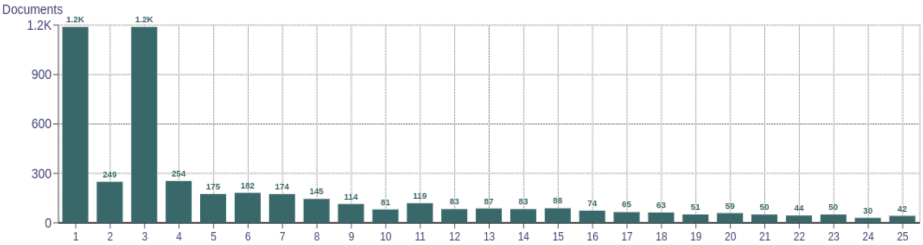
Domain	Docs	% of total
bible.is	1.9K	32.63
wikipedia.org	800	13.98
fakan.ml	785	13.72
jw.org	298	5.21
voabambara.com	233	4.07
rfi.fr	158	2.76
thieme.com	126	2.20
breakeveryyoke.com	106	1.85
wordpress.com	97	1.70
iqna.ir	90	1.57

Top 10 TLDs

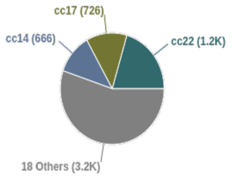
Domain	Docs	% of total
is	1.9K	32.63
org	1.3K	23.25
com	1.1K	18.93
ml	785	13.72
fr	187	3.27
net	150	2.62
ir	103	1.80
co	34	0.59
pl	24	0.42
gov	20	0.35

Documents size (in segments)

<= 25 segments **82.82%** (4.7K documents)
> 25 segments **17.18%** (981 documents)

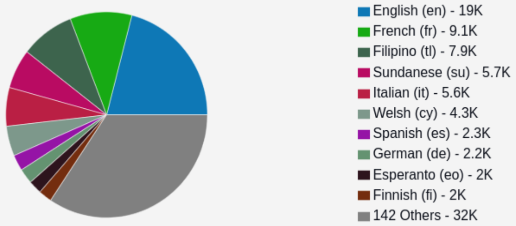


Documents by collection



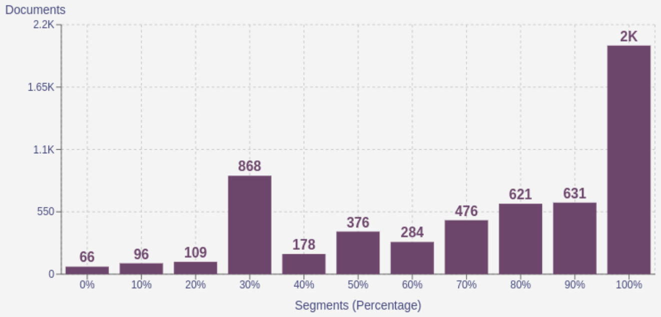
Language Distribution

Number of segments



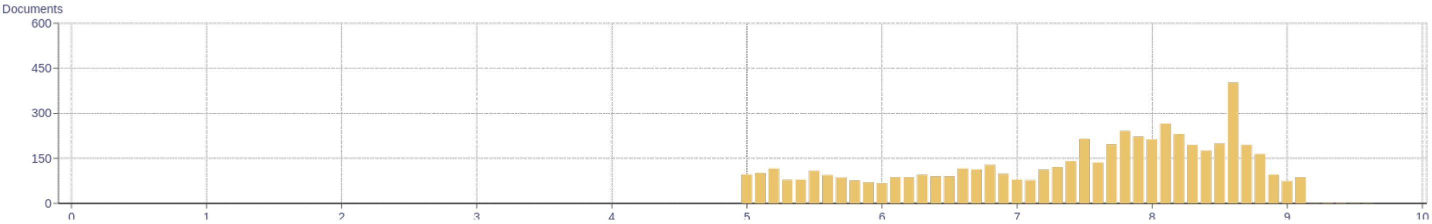
*Bambara (bm) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Bambara (bm) inside documents



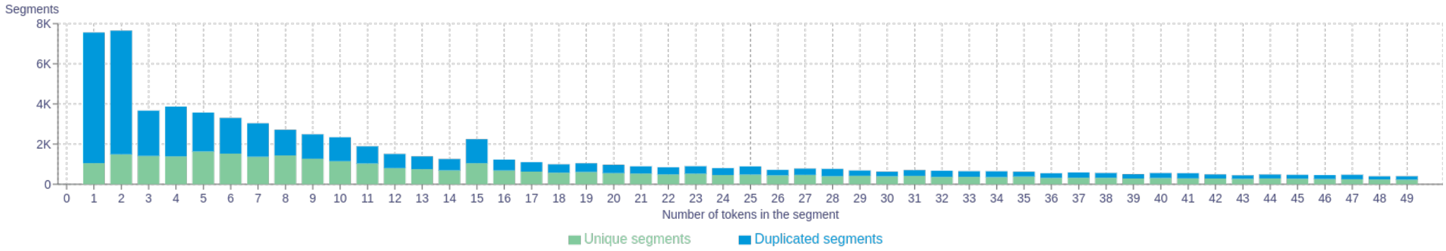
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (5.7K documents)

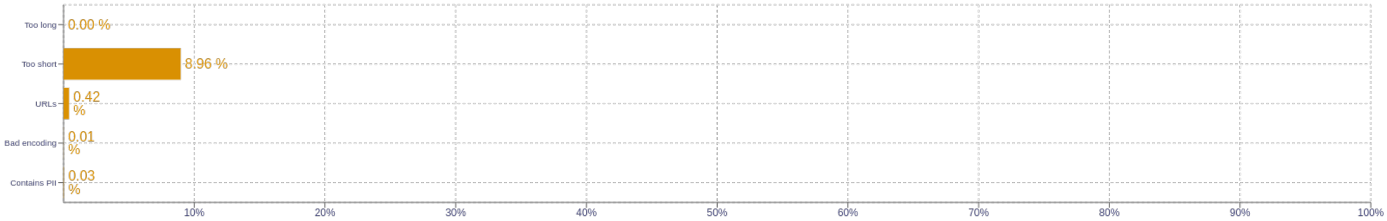


Segment length distribution by token

<= 49 tokens = **32K** segments | **40K** duplicates
> 50 tokens = **20K** segments | **8.6K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	na | 70734 ee | 19956 eee | 18971 bu | 17152 bo | 16720
2	ee eee | 4065 bo bu | 3157 eeeeeeee ee | 2459 na bu | 2354 eee eeeeeeee | 2087
3	eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee | 1396 multi ani mushuleee | 1254 ninvuɣu sheba ban | 1023 bay ka hen | 982 eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee | 859
4	eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee | 1367 eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee | 858 jimou tchi jimou tchi | 610 tchi jimou tchi jimou | 605 bokura wa minna kawaisou | 410
5	eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee | 1338 eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee | 857 tchi jimou tchi jimou tchi | 605 jimou tchi jimou tchi jimou | 605 bo i kɔ atin su | 380

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