HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
bem_Latn.jsonl.tsv	10/4/2024	Bemba (bem)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
6,136	133,538	87,809 (65.76 %)	5.7M	31.03 MB	32,196,810

Top 10 domains

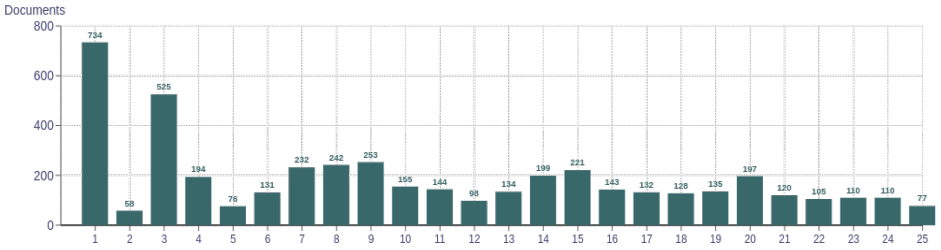
Domain	Docs	% of total
jw.org	4.4K	72.26
bible.is	893	14.55
worldslastchance.com	398	6.49
globalrecordings.net	28	0.46
bibles.org	24	0.39
kingsmanga.net	23	0.37
bible.com	19	0.31
watchtower.org	19	0.31
unicode.org	16	0.26
blogspot.com	12	0.20

Top 10 TLDs

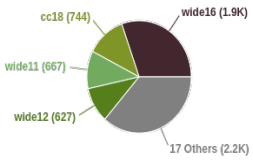
Domain	Docs	% of total
org	4.6K	74.56
is	893	14.55
com	527	8.59
net	69	1.12
cc	10	0.16
info	9	0.15
tv	7	0.11
org.za	6	0.10
co	5	0.08
in	4	0.07

Documents size (in segments)

<= 25 segments **75.83%** (4.7K documents)
> 25 segments **24.17%** (1.5K documents)

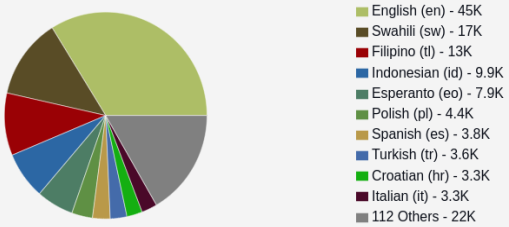


Documents by collection



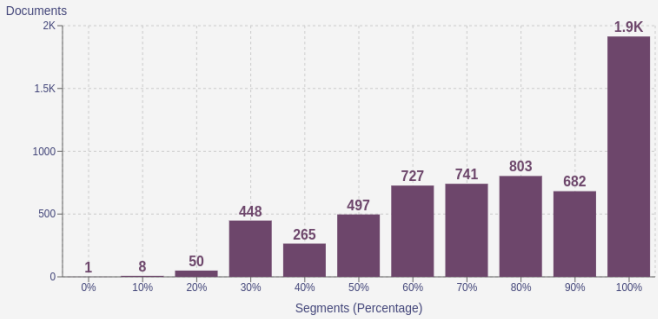
Language Distribution

Number of segments



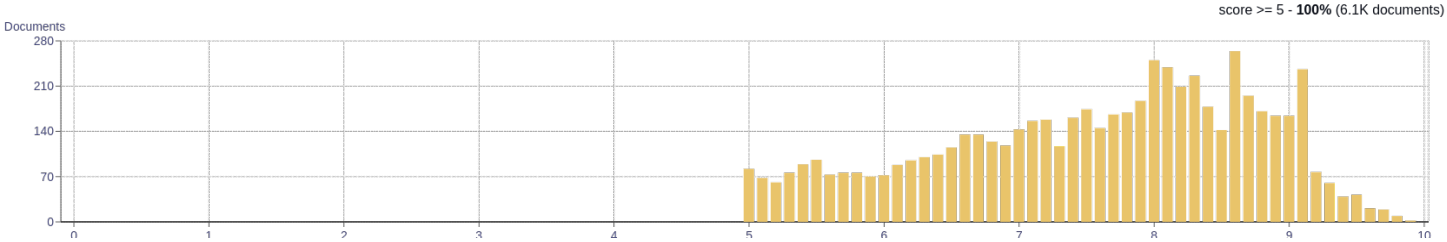
*Bemba (bem) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Bemba (bem) inside documents



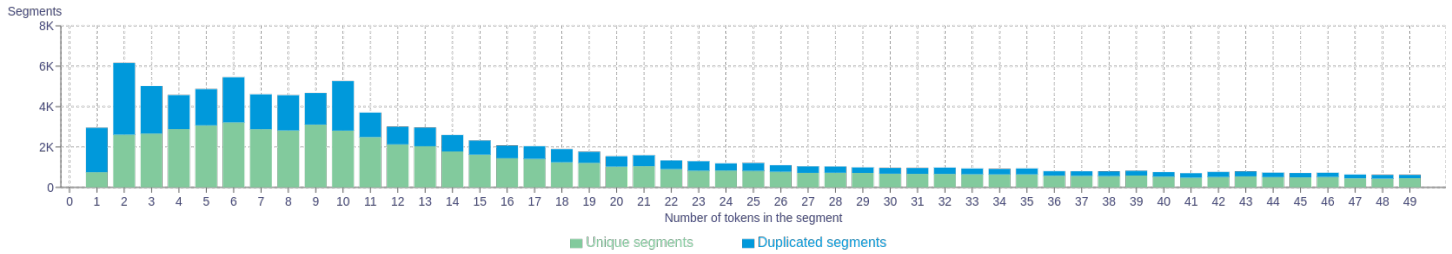
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (6.1K documents)

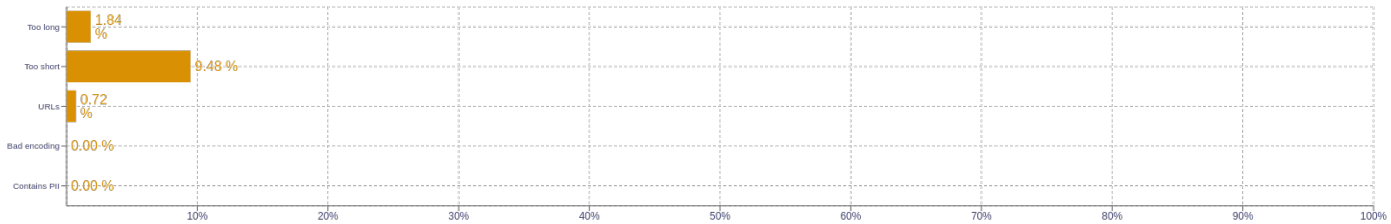


Segment length distribution by token

<= 49 tokens = **62K** segments | **37K** duplicates
> 50 tokens = **35K** segments | **8.7K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	u | 29843 · a | 28044 · yesu | 21515 · i | 19488 · ng | 17335
2	u ng | 2040 · u yesu | 1937 · mukashi banashi | 1724 · furusato saisei | 1717 · mambo a | 1549
3	fumasabomba de fumasabomba | 8710 · nihon no mukashi | 1724 · mukashi banashi episode | 1447 · ore wa shibushibu | 671 · yuusha ni narenakatta | 668
4	nihon no mukashi banashi | 1724 · narenakatta ore wa shibushibu | 671 · bakamonyi ba kwa yehoba | 665 · yuusha ni narenakatta ore | 648 · ore wa shibushibu shuushoku | 604
5	fumasabomba de fumasabomba de fumasabomba | 8708 · nihon no mukashi banashi episode | 1447 · narenakatta ore wa shibushibu shuushoku | 604 · ore wa shibushibu shuushoku o | 358 · wine wine wine wine wine | 234

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