HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
cjk_Latn.jsonl.tsv	12/3/2024	Chokwe (cjk)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,196	36,700	26,104 (71.13 %)	1.2M	7.14 MB	7,395,999

Top 10 domains

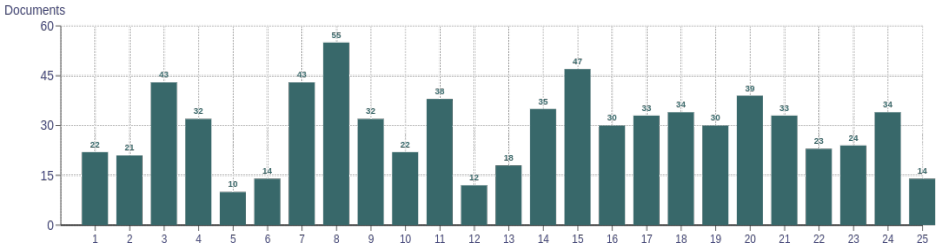
Domain	Docs	% of total
jw.org	1.1K	87.88
globalrecordings.net	20	1.67
unicode.org	18	1.51
bible.is	11	0.92
contafrica.org	9	0.75
watchtower.org	9	0.75
sparkpeople.com	8	0.67
ohchr.org	8	0.67
eatmanga.com	5	0.42
distance2.com	4	0.33

Top 10 TLDs

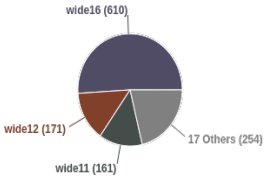
Domain	Docs	% of total
org	1.1K	92.14
com	39	3.26
net	29	2.42
is	11	0.92
in	4	0.33
fr	2	0.17
info	2	0.17
ru	2	0.17
cz	1	0.08
de	1	0.08

Documents size (in segments)

<= 25 segments **61.71%** (738 documents)
> 25 segments **38.29%** (458 documents)

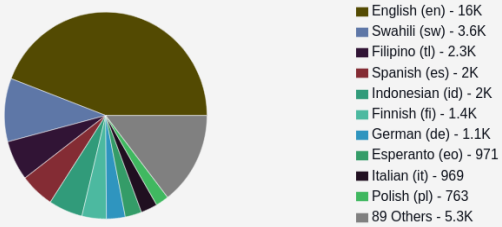


Documents by collection



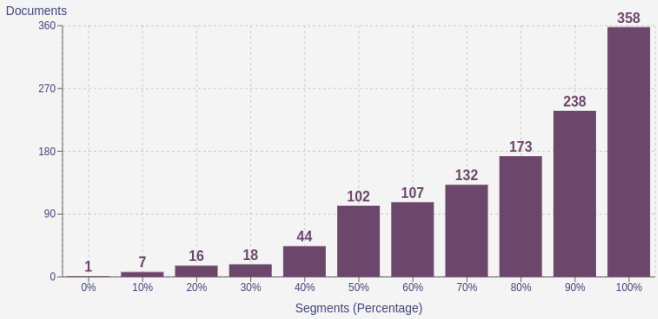
Language Distribution

Number of segments



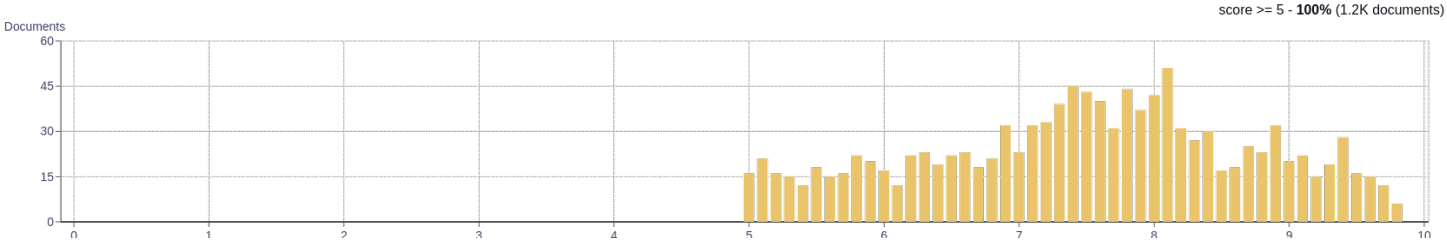
*Chokwe (cjk) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Chokwe (cjk) inside documents



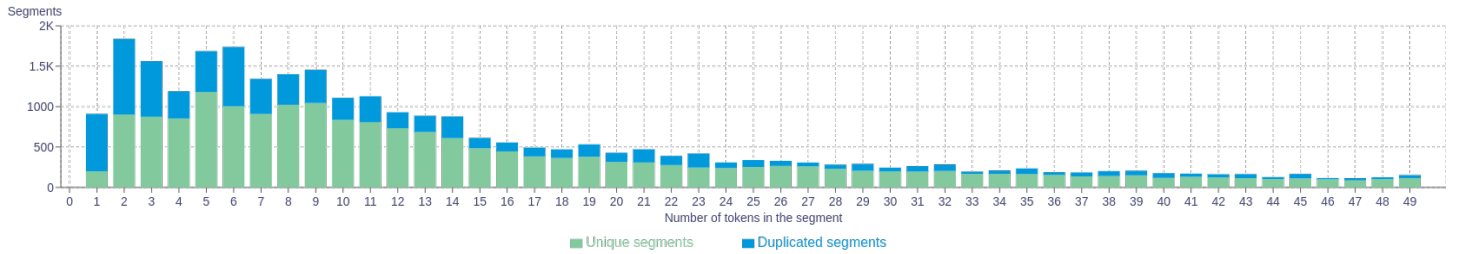
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (1.2K documents)

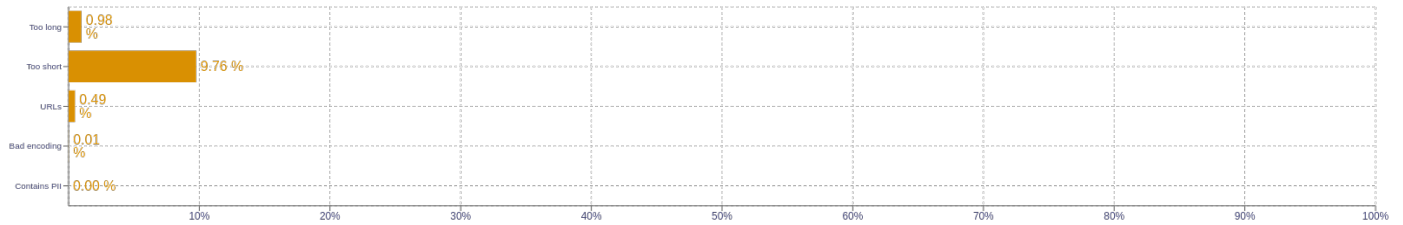


Segment length distribution by token

<= 49 tokens = **19K** segments | **8.9K** duplicates
> 50 tokens = **8.7K** segments | **1.7K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	yehova | 9174 · yesu | 6984 · nawa | 6930 · kaha | 6844 · mu | 6281
2	vyuma muka | 1124 · mwomwo ika | 913 · kaha nawa | 810 · mulong wa | 647 · yiff yiff | 597
3	yiff yiff yiff | 595 · chili chili chili | 396 · haya myaka yosena | 238 · wanangana wa zambi | 219 · kakukk kakukk kakukk | 216
4	yiff yiff yiff yiff | 593 · chili chili chili chili | 395 · kakukk kakukk kakukk kakukk | 213 · world translation of the | 133 · translation of the holy | 133
5	yiff yiff yiff yiff yiff | 591 · chili chili chili chili chili | 394 · kakukk kakukk kakukk kakukk kakukk | 210 · world translation of the holy | 133 · translation of the holy scriptures | 133

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