HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
lua_Latn.jsonl.tsv	11/6/2024	Luba-Lulua (lua)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,083	38,690	27,830 (71.93 %)	1.7M	8.63 MB	8,967,860

Top 10 domains

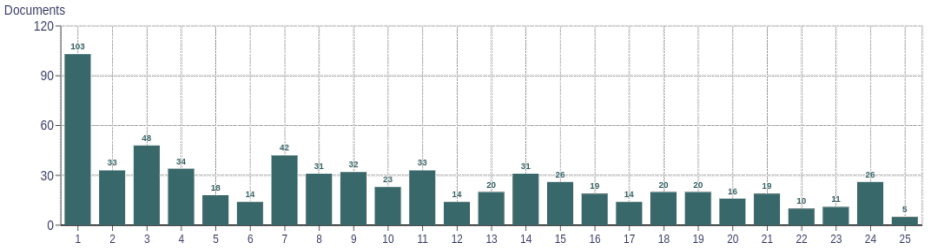
Domain	Docs	% of total
jw.org	871	80.42
biblafrique.net	30	2.77
biblecentre.org	29	2.68
canalblog.com	15	1.39
biblafrique.org	11	1.02
lerythme.ca	8	0.74
mineraiferquebec.com	8	0.74
amq-inc.com	6	0.55
faq-bible.org	6	0.55
distance2.com	6	0.55

Top 10 TLDs

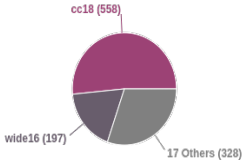
Domain	Docs	% of total
org	945	87.26
com	65	6.00
net	42	3.88
ca	13	1.20
info	4	0.37
ch	3	0.28
xyz	2	0.18
be	2	0.18
tv	2	0.18
co.nz	2	0.18

Documents size (in segments)

<= 25 segments **61.13%** (662 documents)
> 25 segments **38.87%** (420 documents)

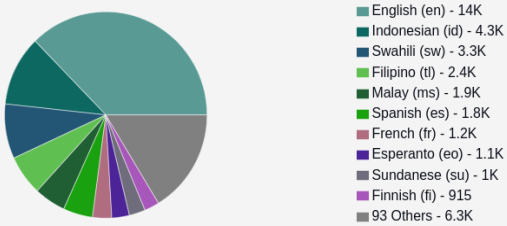


Documents by collection



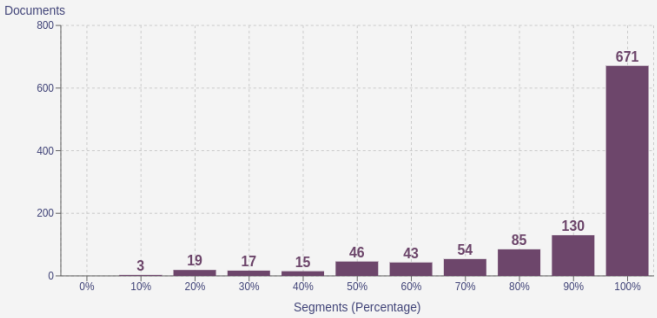
Language Distribution

Number of segments



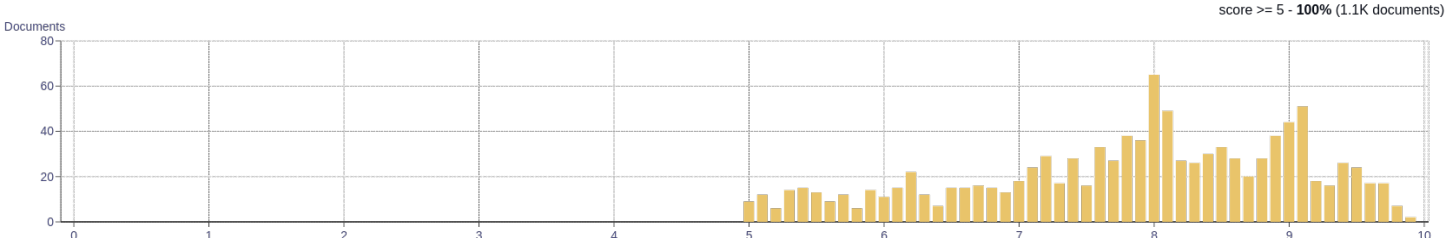
*Luba-Lulua (lua) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Luba-Lulua (lua) inside documents



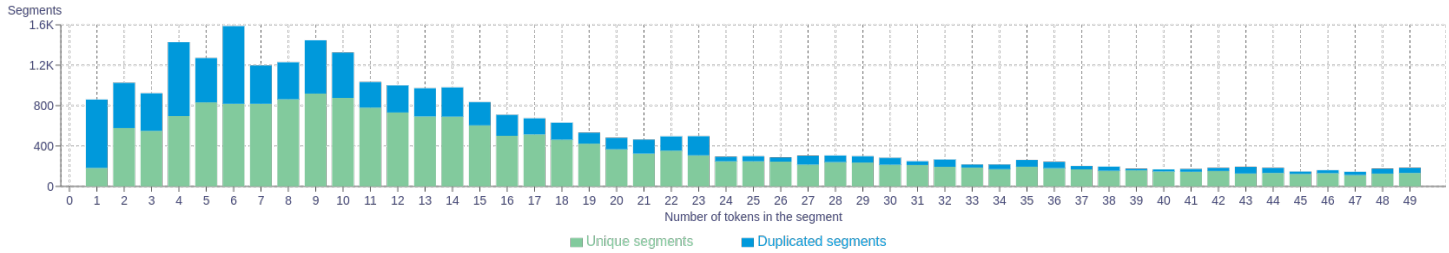
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (1.1K documents)

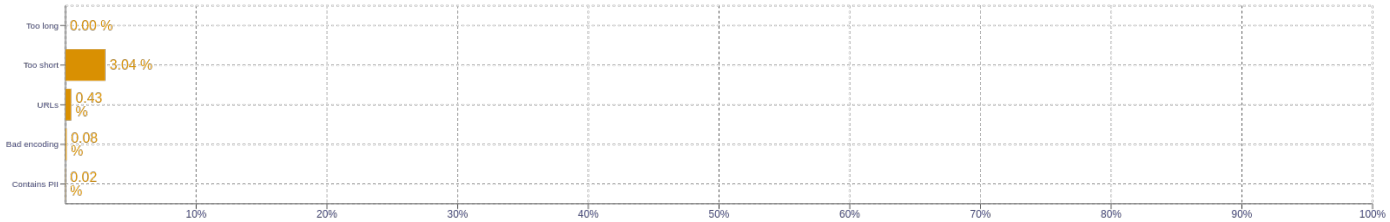


Segment length distribution by token

<= 49 tokens = **19K** segments | **8.9K** duplicates
> 50 tokens = **11K** segments | **1.9K** duplicates



Segment noise distribution



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