HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
lin_Latn.jsonl.tsv	11/11/2024	Lingala (ln)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
7,588	200,341	111,837 (55.82 %)	6.6M	31.86 MB	32,731,201

Top 10 domains

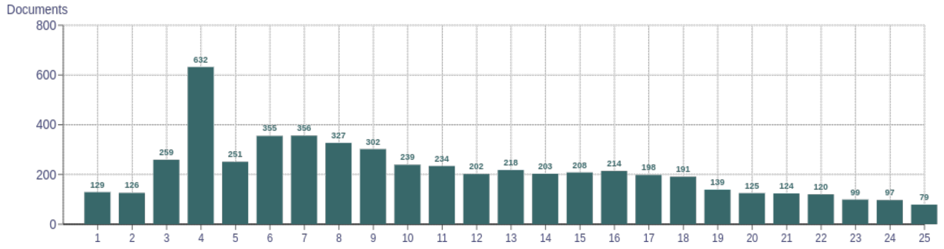
Domain	Docs	% of total
jw.org	2.1K	27.19
wikipedia.org	1.1K	14.55
voalingala.com	751	9.90
mbokamosika.com	602	7.93
lds.org	262	3.45
skyrock.com	232	3.06
senemongaba.com	225	2.97
migraweb.ch	120	1.58
congomikili.com	117	1.54
voiceofcongo.net	110	1.45

Top 10 TLDs

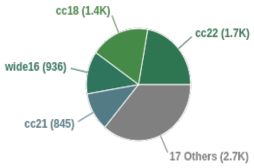
Domain	Docs	% of total
org	3.7K	48.30
com	2.8K	36.64
net	428	5.64
ch	132	1.74
cat	85	1.12
fr	81	1.07
info	76	1.00
gov	45	0.59
nl	27	0.36
ru	22	0.29

Documents size (in segments)

<= 25 segments **71.52%** (5.4K documents)
> 25 segments **28.48%** (2.2K documents)

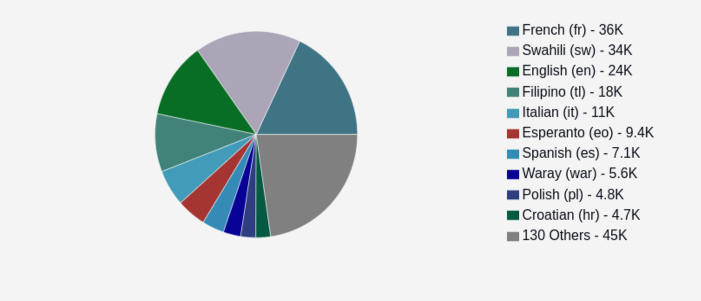


Documents by collection



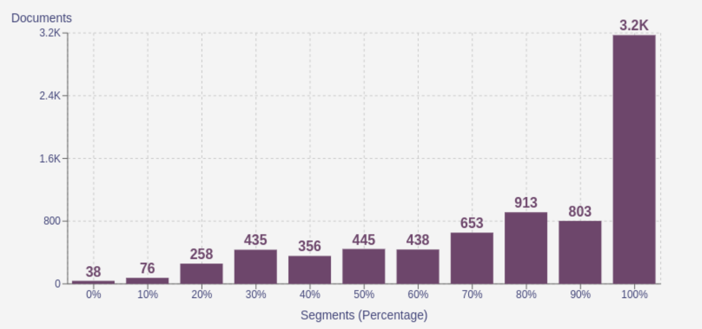
Language Distribution

Number of segments



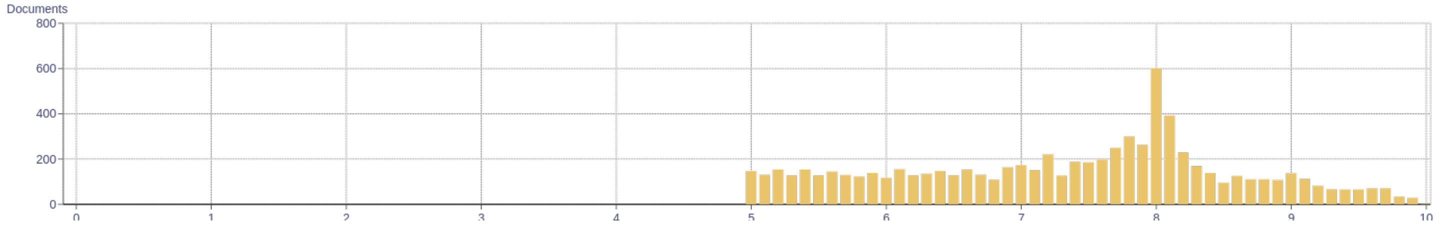
*Lingala (ln) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Lingala (ln) inside documents



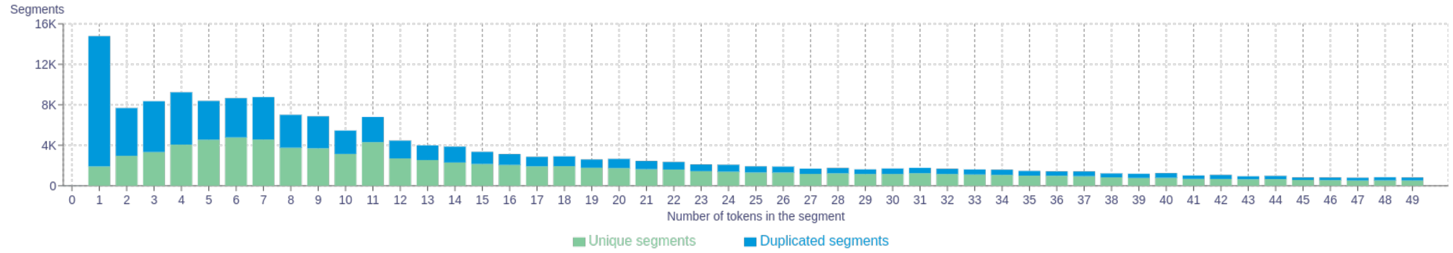
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (7.6K documents)

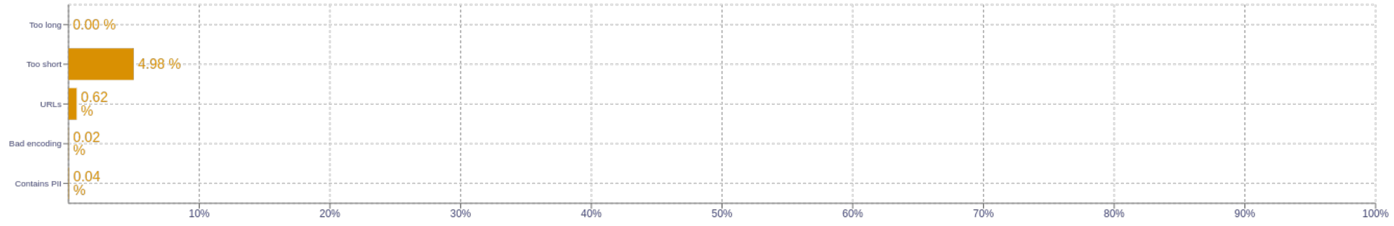


Segment length distribution by token

<= 49 tokens = **89K** segments | **75K** duplicates
> 50 tokens = **36K** segments | **13K** 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