HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
sna_Latn.jsonl.tsv	11/28/2024	Shona (sn)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
61,076	1,201,679	864,345 (71.93 %)	29M	183.12 MB	191,477,676

Top 10 domains

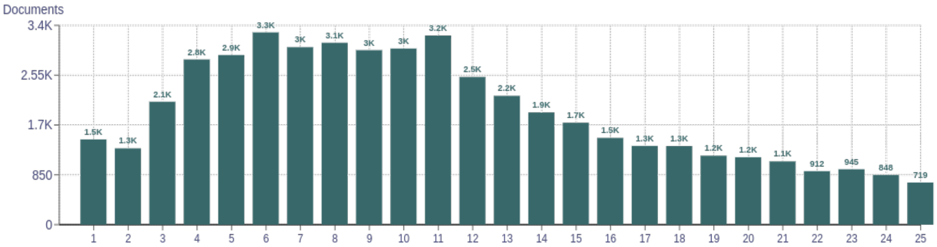
Domain	Docs	% of total
voashona.com	6.7K	11.04
wikipedia.org	5.5K	8.95
jw.org	5K	8.23
linuxadictos.com	2.3K	3.84
eturbonews.com	1.6K	2.66
kwayedza.co.zw	1.6K	2.57
martech.zone	1K	1.64
actualidadiphone.com	857	1.40
zimkatorike.com	726	1.19
masasieharare.com	642	1.05

Top 10 TLDs

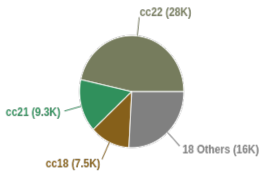
Domain	Docs	% of total
com	41K	66.65
org	12K	20.40
co.zw	2.3K	3.74
zone	1K	1.64
net	998	1.63
africa	374	0.61
fr	214	0.35
co.za	212	0.35
ru	194	0.32
org.uk	170	0.28

Documents size (in segments)

<= 25 segments **79.46%** (49K documents)
> 25 segments **20.54%** (13K documents)

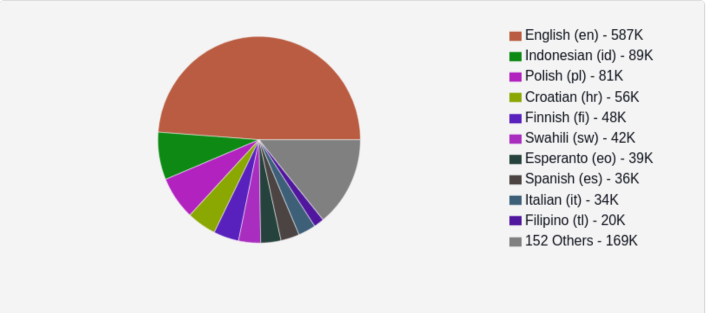


Documents by collection



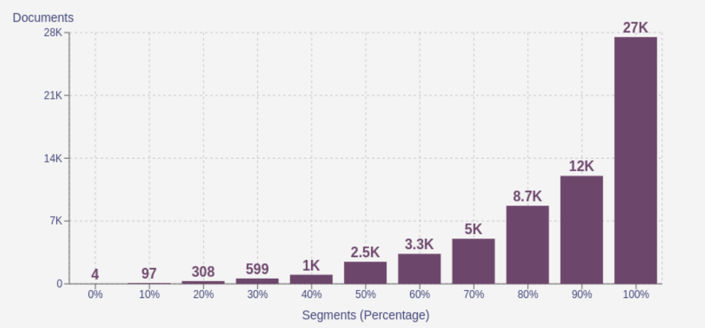
Language Distribution

Number of segments



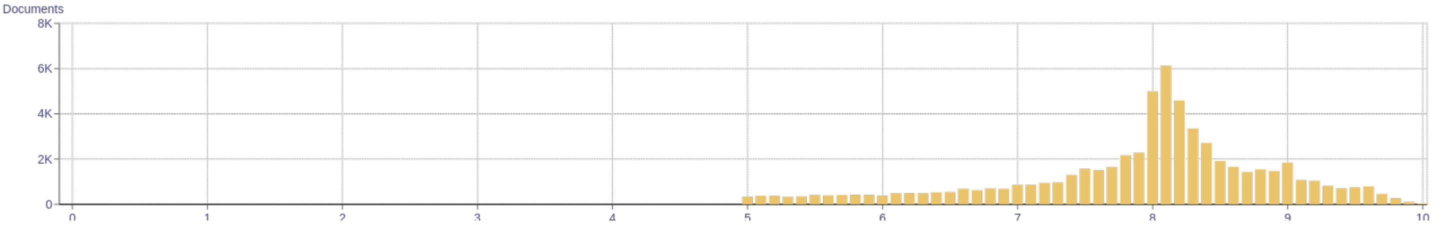
*Shona (sn) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Shona (sn) inside documents



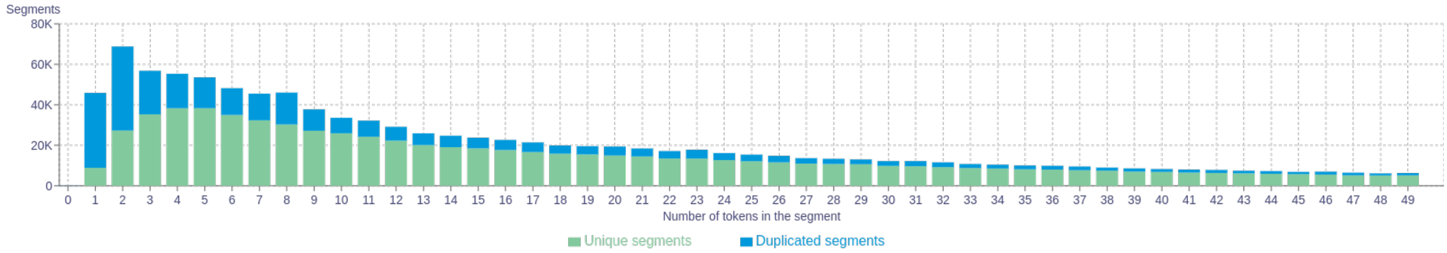
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (61K documents)

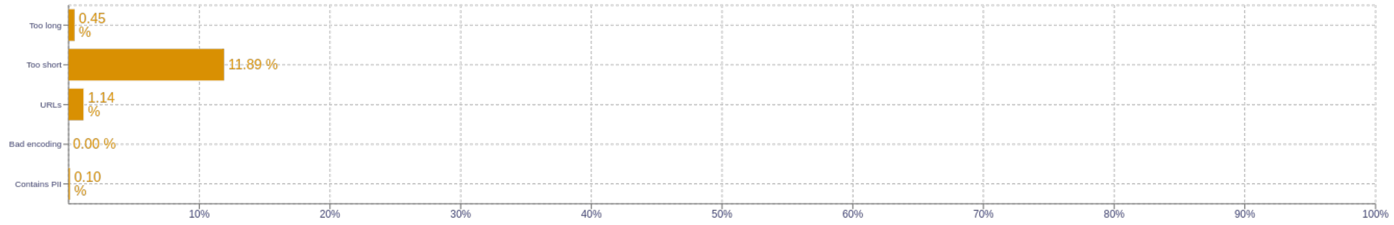


Segment length distribution by token

<= 49 tokens = **737K** segments | **309K** duplicates
> 50 tokens = **156K** segments | **29K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	kana | 282828 iyo | 138946 asi | 102289 kubva | 86647 iri | 73004
2	chirp chirp | 15672 zviri nyore | 6956 edit source | 6669 makumi maviri | 6263 imwe chete | 4739
3	chirp chirp chirp | 15667 kana iwe uchida | 4547 uchinge uchinge uchinge | 3636 kana iwe uri | 2168 panguva imwe chete | 1991
4	chirp chirp chirp chirp | 15664 uchinge uchinge uchinge uchinge | 3488 kuverenga nyaya ino mumutauro | 1379 here kuverenga nyaya ino | 1379 yenyika itsva yemagwaro matsvene | 1355
5	chirp chirp chirp chirp chirp | 15661 uchinge uchinge uchinge uchinge uchinge | 3347 ungada here kuverenga nyaya ino | 1379 kuverenga nyaya ino mumutauro we | 1379 here kuverenga nyaya ino mumutauro | 1379

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