HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
tso_Latn.jsonl.tsv	9/19/2024	Tsonga (ts)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
11,008	221,245	136,723 (61.80 %)	10M	47.58 MB	49,075,139

Top 10 domains

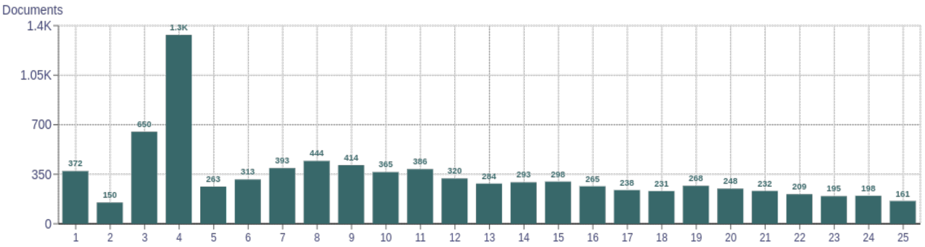
Domain	Docs	% of total
jw.org	6.5K	58.92
biblesa.co.za	1K	9.49
wikipedia.org	999	9.08
bible.is	653	5.93
southafrica.co.za	427	3.88
vivmag.co.za	243	2.21
rivoni.org	54	0.49
munghanalonenefm.co.za	45	0.41
myconstitution.co.za	40	0.36
matimunews.co.za	38	0.35

Top 10 TLDs

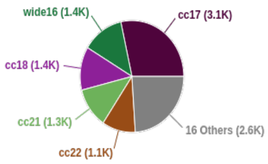
Domain	Docs	% of total
org	7.8K	70.60
co.za	2K	18.10
is	653	5.93
com	252	2.29
net	99	0.90
gov.za	82	0.74
org.za	44	0.40
ac.za	33	0.30
africa	13	0.12
ch	11	0.10

Documents size (in segments)

<= 25 segments **77.43%** (8.5K documents)
> 25 segments **22.57%** (2.5K documents)

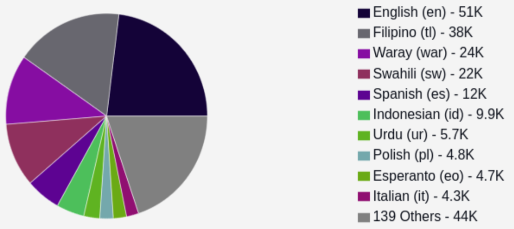


Documents by collection



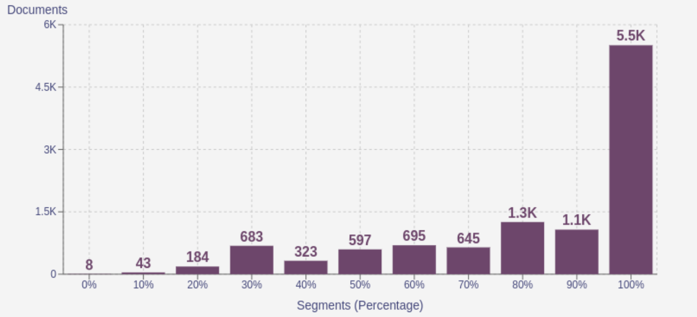
Language Distribution

Number of segments



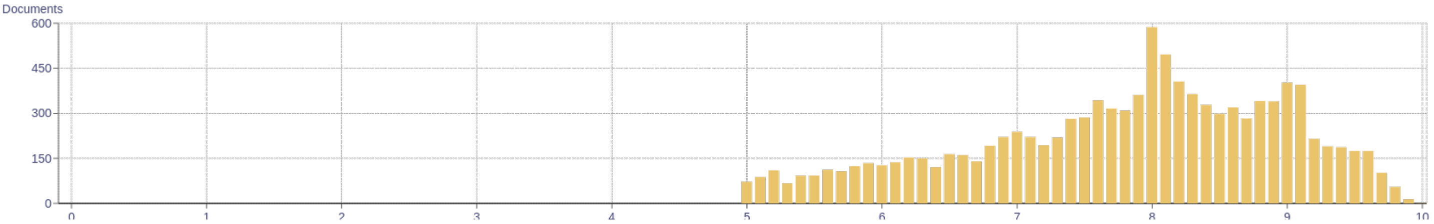
*Tsonga (ts) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Tsonga (ts) inside documents



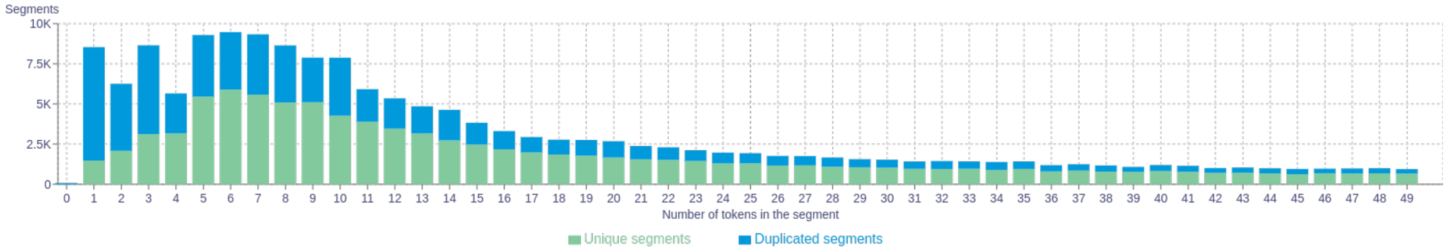
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (11K documents)

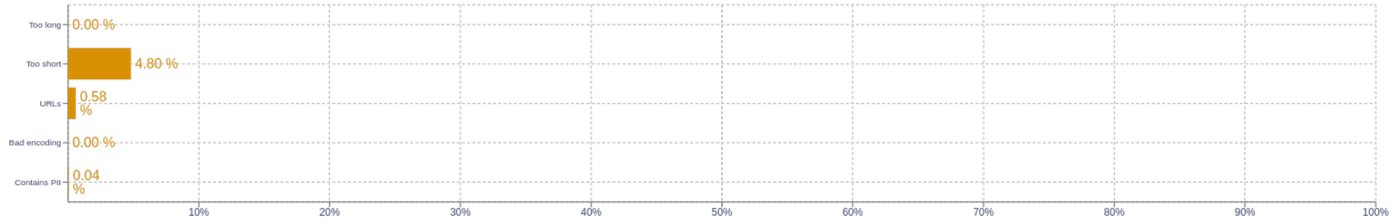


Segment length distribution by token

<= 49 tokens = **94K** segments | **68K** duplicates
> 50 tokens = **60K** segments | **17K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ni | 165530 vha | 63263 ha | 59393 wana | 52896 n | 46262
2	ndlela leyi | 10450 ha yona | 8602 ha yini | 6061 wana ni | 5889 vanhu lava | 5636
3	timbhoni ta yehovha | 3870 wana ni un | 2301 vanhu vo tala | 2125 bya misava leyintshwa | 1677 vuhundzuluxeri bya misava | 1671
4	vuhundzuluxeri bya misava leyintshwa | 1664 bya misava leyintshwa bya | 1294 misava leyintshwa bya matsalwa | 1271 leyintshwa bya matsalwa yo | 1250 bya matsalwa yo kwetsima | 1250
5	vuhundzuluxeri bya misava leyintshwa bya | 1294 bya misava leyintshwa bya matsalwa | 1271 misava leyintshwa bya matsalwa yo | 1250 leyintshwa bya matsalwa yo kwetsima | 1250 tsakela ku hlaya xihloko lexi | 1188

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