HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
tsn_Latn.jsonl.tsv	9/20/2024	Tswana (tn)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
6,050	132,173	77,616 (58.72 %)	6.1M	26.44 MB	27,545,230

Top 10 domains

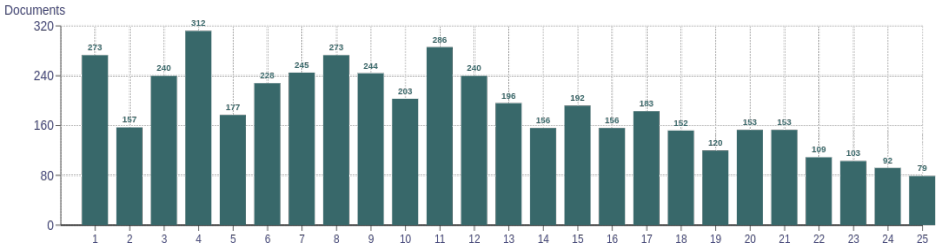
Domain	Docs	% of total
jw.org	2.4K	39.95
wikipedia.org	1.2K	19.49
biblesa.co.za	530	8.76
southafrica.co.za	247	4.08
gov.bw	145	2.40
sciencegraph.net	145	2.40
nwu.ac.za	137	2.26
dikgang24.news	76	1.26
oxforddictionaries.com	70	1.16
mmegi.bw	44	0.73

Top 10 TLDs

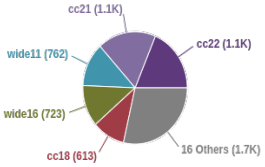
Domain	Docs	% of total
org	3.8K	62.86
co.za	976	16.13
com	390	6.45
net	190	3.14
bw	190	3.14
ac.za	151	2.50
news	76	1.26
support	38	0.63
org.za	37	0.61
gov.za	37	0.61

Documents size (in segments)

<= 25 segments **78.05%** (4.7K documents)
> 25 segments **21.95%** (1.3K documents)

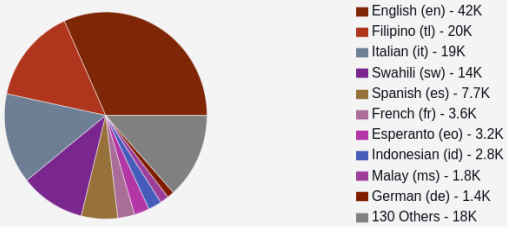


Documents by collection



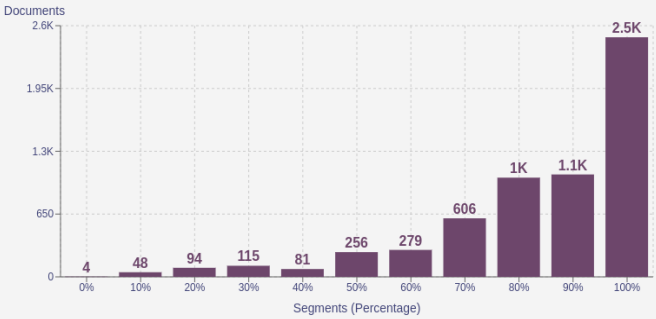
Language Distribution

Number of segments



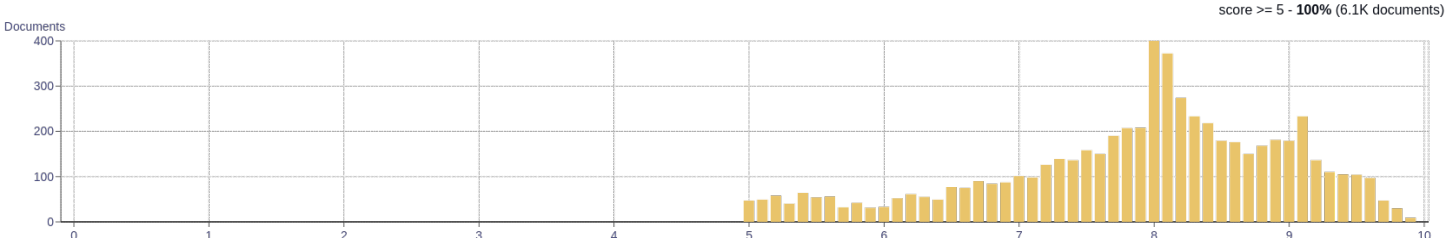
*Tswana (tn) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Tswana (tn) inside documents



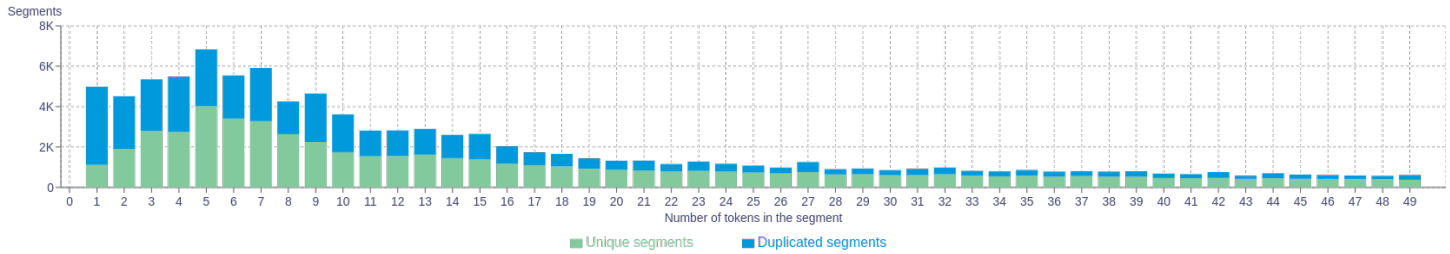
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (6.1K documents)

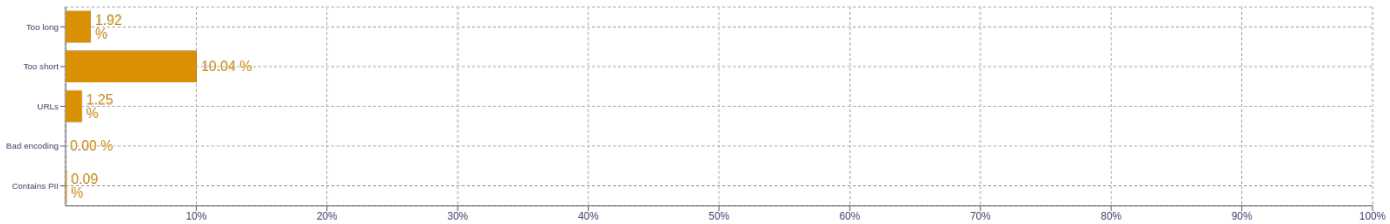


Segment length distribution by token

<= 49 tokens = **55K** segments | **43K** duplicates
> 50 tokens = **34K** segments | **12K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	a | 225859 · ya | 123090 · wa | 52375 · tse | 46028 · mme | 42568
2	a a | 10824 · boing boing | 7724 · jo bo | 5198 · neng a | 4639 · tse dingwe | 4146
3	boing boing boing | 7723 · a ne a | 3938 · a neng a | 3122 · a bo a | 2156 · a ba a | 1072
4	boing boing boing boing | 7722 · jesu o ne a | 1598 · basupi ba ga jehofa | 1483 · jehofa o ne a | 1083 · a se ka a | 1034
5	boing boing boing boing boing | 7721 · jaana o meri jaana film | 505 · song jaana o meri jaana | 504 · tsuki ga michibiku isekai douchuu | 471 · thanolo ya lefatshe le lesha | 391

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