HPLT Analytics report

General overview

Corpus	Analytics date	Language
sot_Latn.jsonl.tsv	9/22/2024	Southern Sotho (st)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
43,917	1,085,450	798,020 (73.52 %)	36M	163.42 MB	170,450,093

Top 10 domains

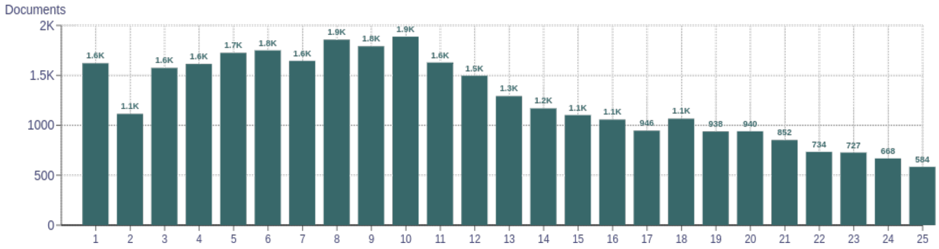
Domain	Docs	% of total
jw.org	1.8K	4.14
eturbonews.com	1.7K	3.98
martech.zone	1.1K	2.41
comme-un-pro.fr	695	1.58
educationbro.com	512	1.17
actualidadiphone.com	481	1.10
bibles.org	447	1.02
actualidadgadget.com	432	0.98
actualidadliteratura.com	415	0.94
hombresconestilo.com	350	0.80

Top 10 TLDs

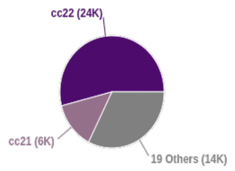
Domain	Docs	% of total
com	30K	68.03
org	4.8K	10.91
zone	1.1K	2.41
net	1.1K	2.40
info	1K	2.34
co.za	818	1.86
fr	724	1.65
org.za	474	1.08
ru	393	0.89
es	305	0.69

Documents size (in segments)

<= 25 segments **72.36%** (32K documents)
> 25 segments **27.64%** (12K documents)

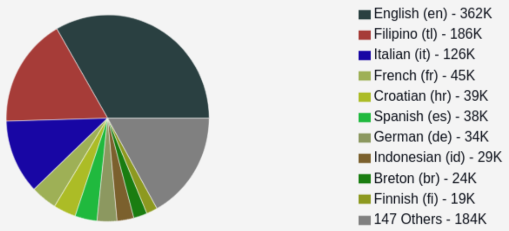


Documents by collection



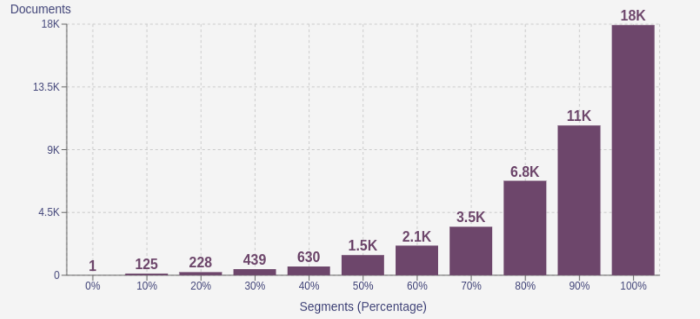
Language Distribution

Number of segments



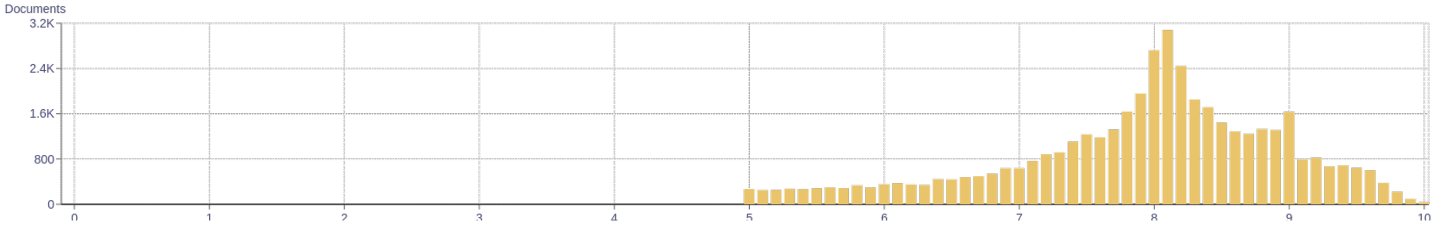
*Southern Sotho (st) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Southern Sotho (st) inside documents



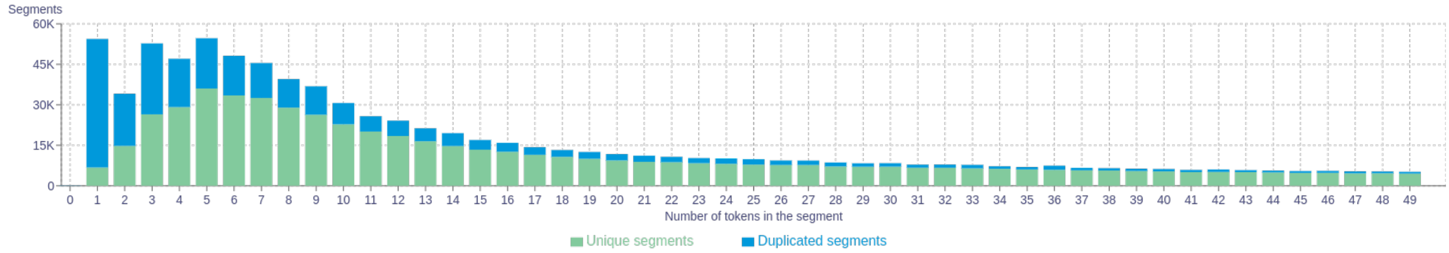
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (44K documents)

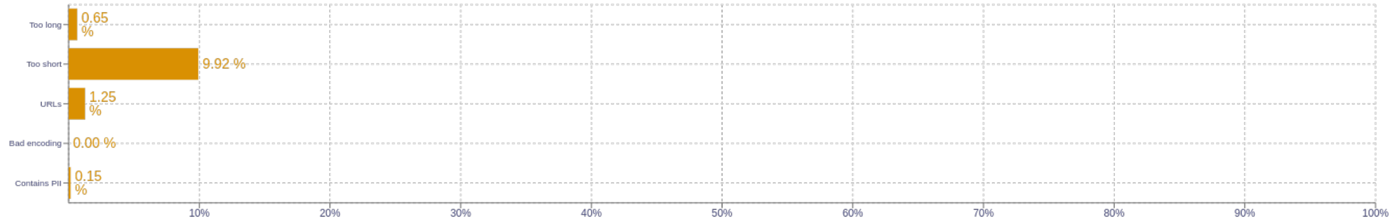


Segment length distribution by token

<= 49 tokens = **578K** segments | **259K** duplicates
> 50 tokens = **249K** segments | **29K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	u | 273841 na | 202107 tla | 154364 kapa | 152701 bakeng | 129172
2	haeba u | 25463 u tla | 14698 hona joale | 12952 boleng bo | 11317 na u | 10826
3	nako e telele | 9548 efe kapa efe | 8630 kantle ho naha | 6413 bophelo bo botle | 5845 motho e mong | 5838
4	leha ho le joalo | 16266 mong le e mong | 9449 molemo ka ho fetisisa | 7918 leha e le efe | 6768 u se ke ua | 6086
5	ding ding ding ding ding | 2725 sebaka sa hau sa marang | 1064 etsa bonnete ba hore u | 1009 boitsebiso leha e le bofe | 991 tabeng ena bo fumaneha amazon | 948

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