

General overview

Corpus	Analytics date	Language
xho_Latn.jsonl.tsv	11/11/2024	Xhosa (xh)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
63,091	1,820,800	1,056,230 (58.01 %)	39M	245.75 MB	256,905,047

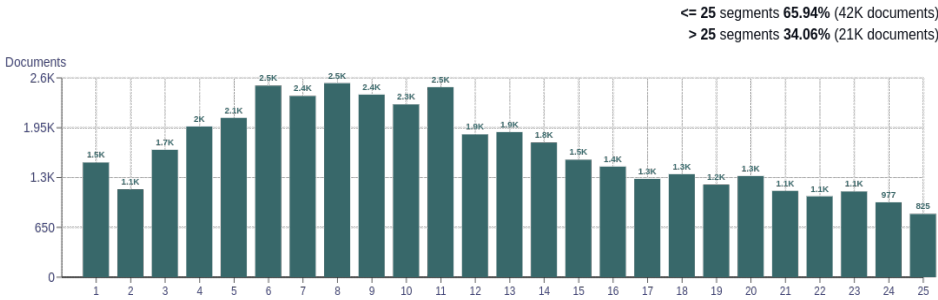
Top 10 domains

Domain	Docs	% of total
airbnb.com	7.1K	11.22
jw.org	4.5K	7.13
isolezwelesixhosa.co.za	4.3K	6.87
wordplanet.org	2.6K	4.10
wikipedia.org	1.4K	2.17
creativosonline.org	1.3K	2.10
martech.zone	1.1K	1.70
actualidadgadget.com	766	1.21
recetin.com	725	1.15
wordproject.org	660	1.05

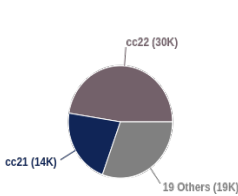
Top 10 TLDs

Domain	Docs	% of total
com	36K	57.80
org	12K	19.41
co.za	5.8K	9.22
zone	1.1K	1.70
net	857	1.36
top	754	1.20
pt	710	1.13
info	406	0.64
date	349	0.55
org.za	330	0.52

Documents size (in segments)

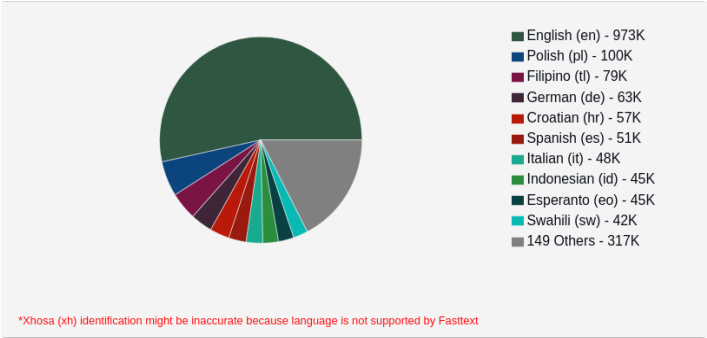


Documents by collection

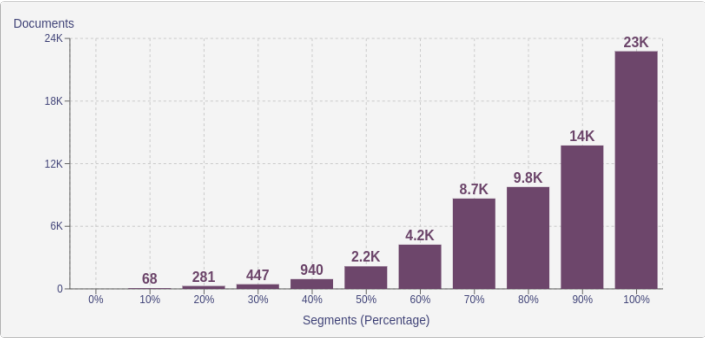


Language Distribution

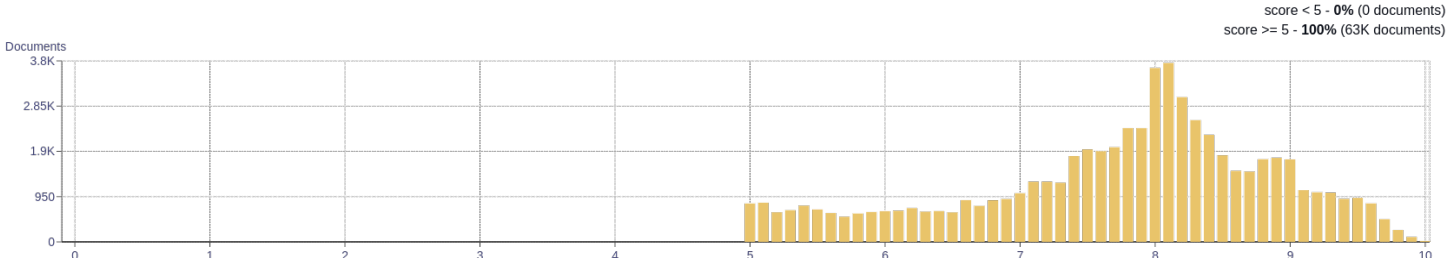
Number of segments



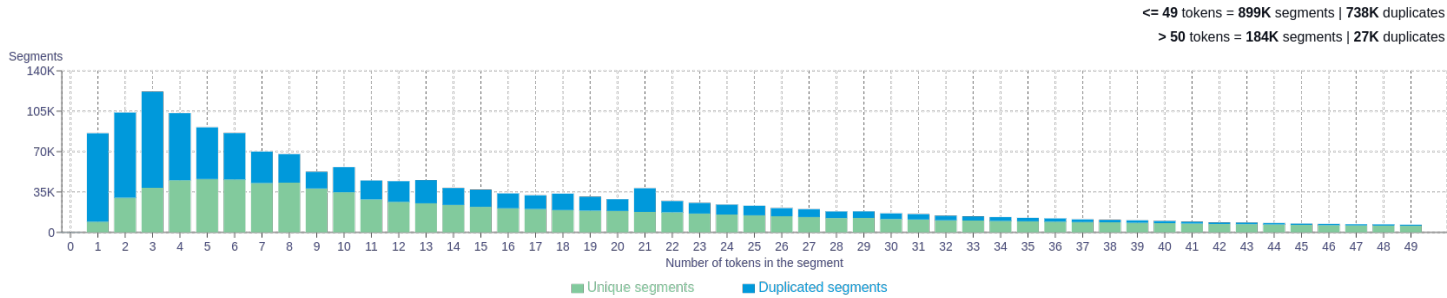
Percentage of segments in Xhosa (xh) inside documents



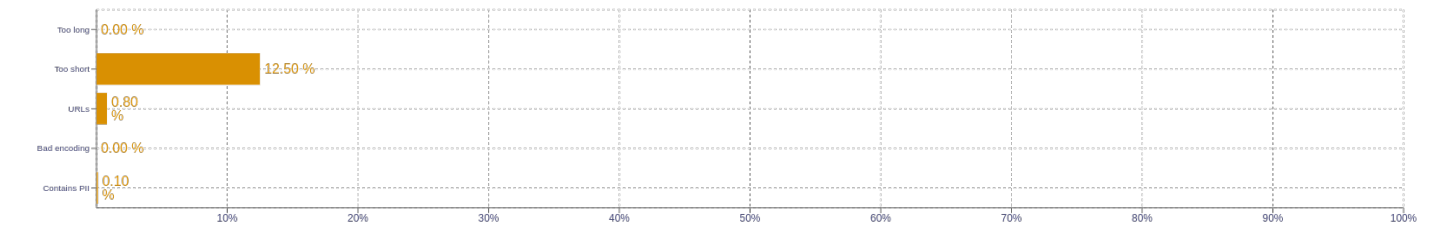
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of types (uniques)/number of tokens*, after removing punctuation (<https://www.stinfo.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>, , , etc.) replaced by newlines.

Language distribution

Language identified with FastSpell (<https://github.com/mbanoni/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/pabiop16n/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>