HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
af_1.jsonl.tsv	3/21/2024	Afrikaans (af)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
747,229	84,701,482	19,402,074 (22.91 %)	1B	4.99 GB	

Top 10 domains

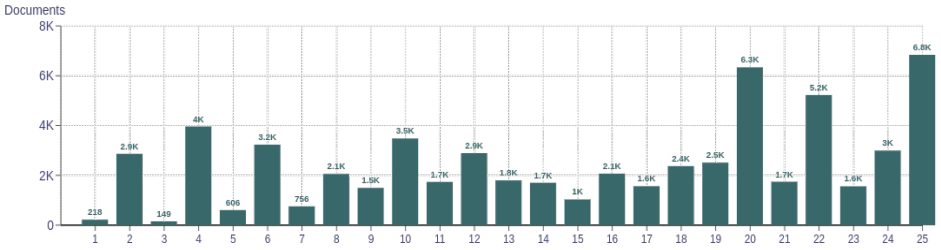
Domain	Docs	% of total
diebuchsuche.com	408K	54.62
journals.co.za	38K	5.14
wikipedia.org	12K	1.65
maroelamedia.co.za	10K	1.40
netwerk24.com	7.1K	0.95
praag.co.za	6.3K	0.85
kosmos.com.na	5.6K	0.75
watkykjy.co.za	4.1K	0.54
roekeloos.co.za	4K	0.53
landbou.com	3.6K	0.48

Top 10 TLDs

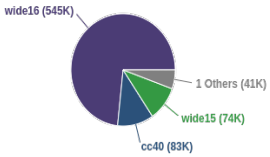
Domain	Docs	% of total
com	509K	68.17
co.za	140K	18.80
org	29K	3.94
net	8.8K	1.18
org.za	6.2K	0.84
com.na	6.1K	0.81
info	4.4K	0.58
nl	3.8K	0.51
ac.za	3.6K	0.48
ca	2.5K	0.34

Documents size (in segments)

<= 25 segments **8.23%** (61K documents)
> 25 segments **91.77%** (682K documents)

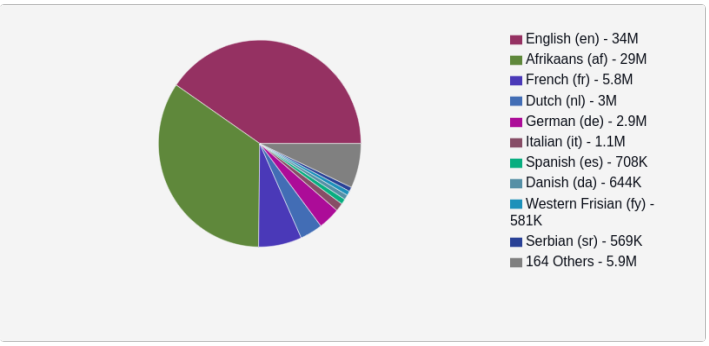


Documents by collection

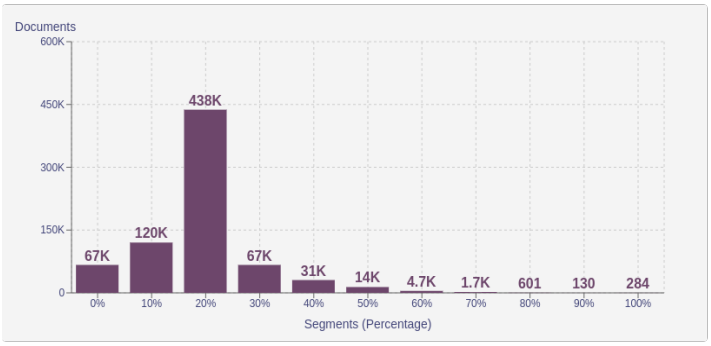


Language Distribution

Number of segments

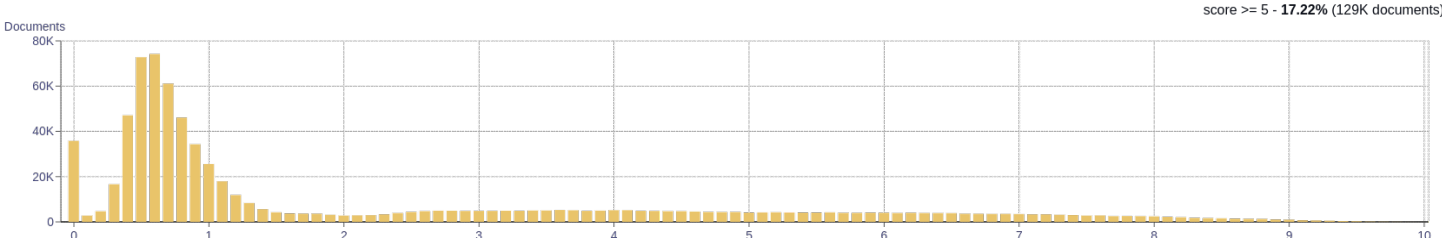


Percentage of segments in Afrikaans (af) inside documents



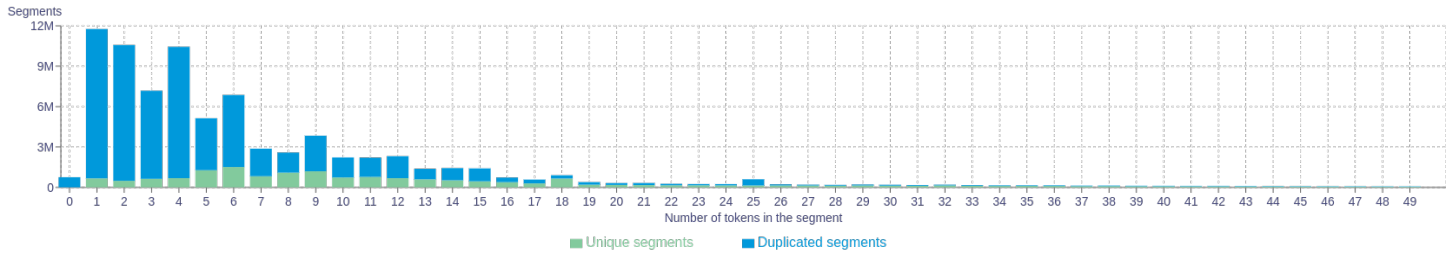
Distribution of documents by document score

score < 5 - **82.78%** (619K documents)
score >= 5 - **17.22%** (129K documents)

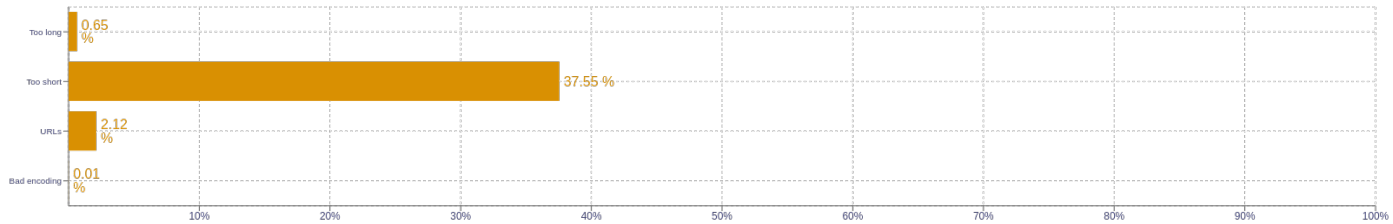


Segment length distribution by token

<= 49 tokens = **17M** segments | **64M** duplicates
> 50 tokens = **3.6M** segments | **1.3M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	the | 13700283 · of | 12108285 · and | 8106124 · to | 5890313 · a | 4994182
2	of the | 2027007 · meer besonderhede | 1355889 · alternatiewe skryfwyses | 1352488 · kyk boek | 1346521 · stoor boek | 1346517
3	to my favourites | 410387 · made by freepik | 408232 · freepik from www.flaticon.com | 408232 · www.flaticon.com is licensed | 408231 · licensed by cc | 408231
4	add to my favourites | 410386 · made by freepik from | 408232 · icons made by freepik | 408232 · from www.flaticon.com is licensed | 408231 · oor die boek soek | 404630
5	made by freepik from www.flaticon.com | 408232 · icons made by freepik from | 408232 · www.flaticon.com is licensed by cc | 408231 · freepik from www.flaticon.com is licensed | 408231 · united kingdom australia new zealand | 387419

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