

General overview

Corpus	Analytics date	Language
afr_Latn.json.tsv	9/6/2024	Afrikaans (af)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,457,165	37,737,319	18,802,427 (49.82 %)	1.2B	5.56 GB	5,910,906,748

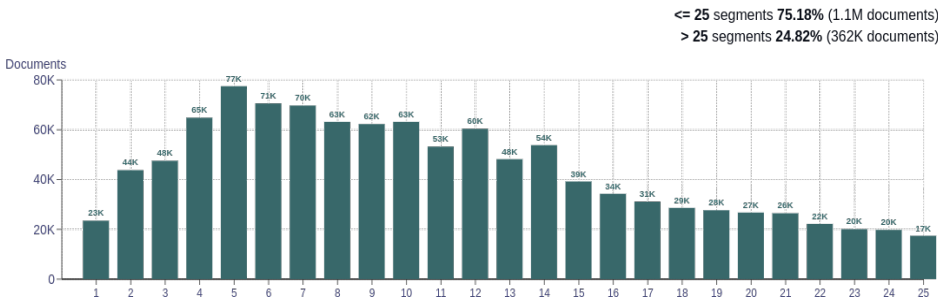
Top 10 domains

Domain	Docs	% of total
wikipedia.org	164K	11.23
maroelamedia.co.za	68K	4.64
netwerk24.com	42K	2.87
landbou.com	35K	2.38
praag.co.za	33K	2.24
wordpress.com	32K	2.19
litnet.co.za	32K	2.17
sarie.com	28K	1.90
software.net	20K	1.40
androware.net	20K	1.39

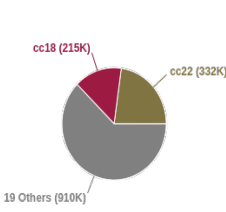
Top 10 TLDs

Domain	Docs	% of total
com	529K	36.33
co.za	451K	30.96
org	239K	16.38
net	69K	4.74
org.za	30K	2.06
ac.za	28K	1.90
com.na	19K	1.27
ca	11K	0.76
info	9.3K	0.64
pt	4K	0.28

Documents size (in segments)

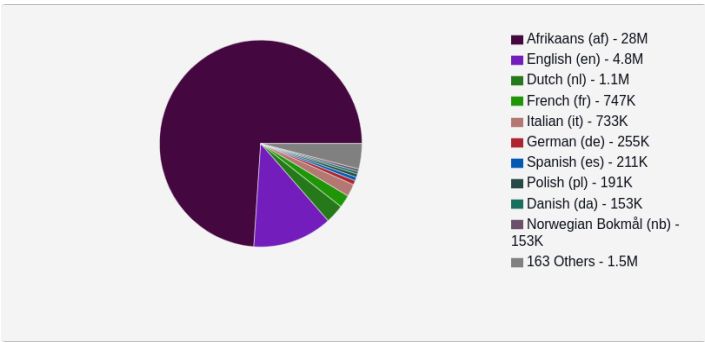


Documents by collection

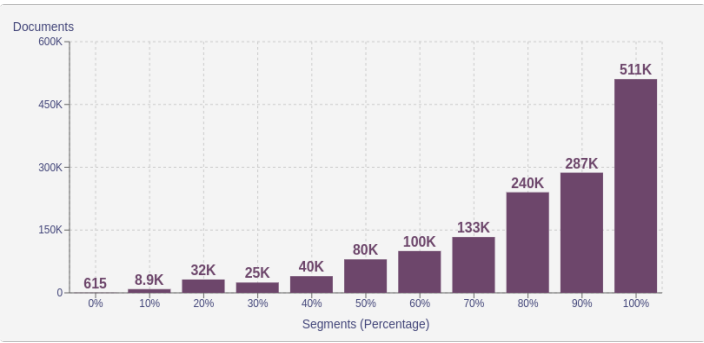


Language Distribution

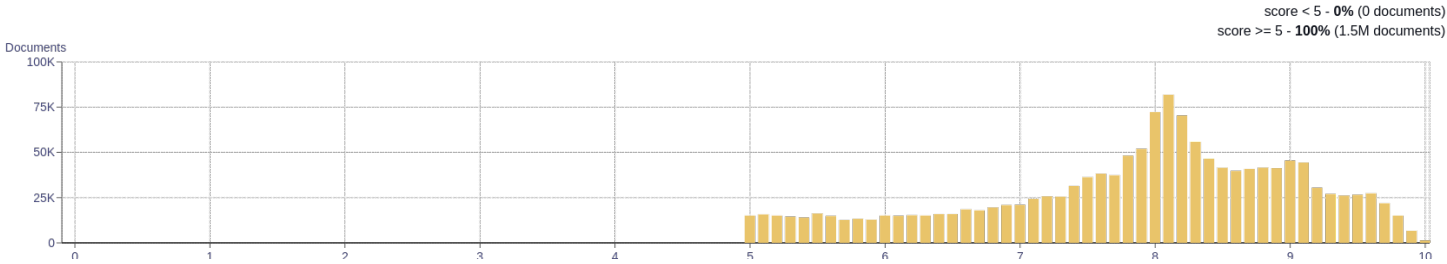
Number of segments



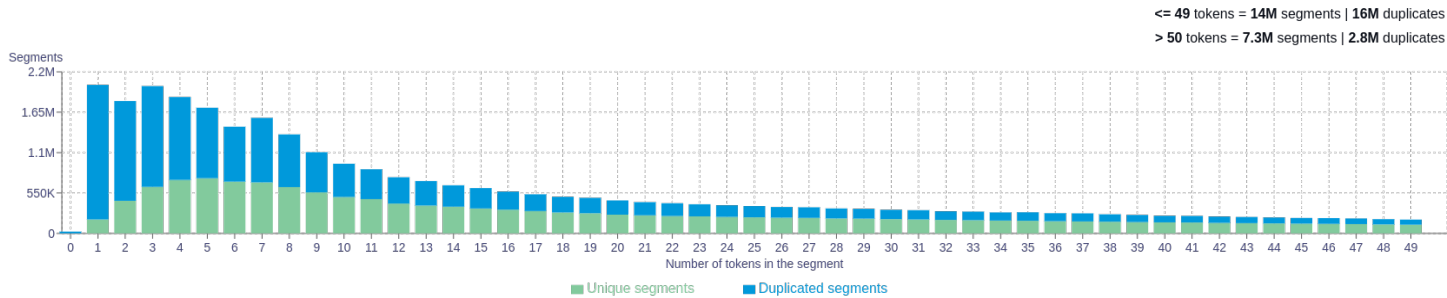
Percentage of segments in Afrikaans (af) inside documents



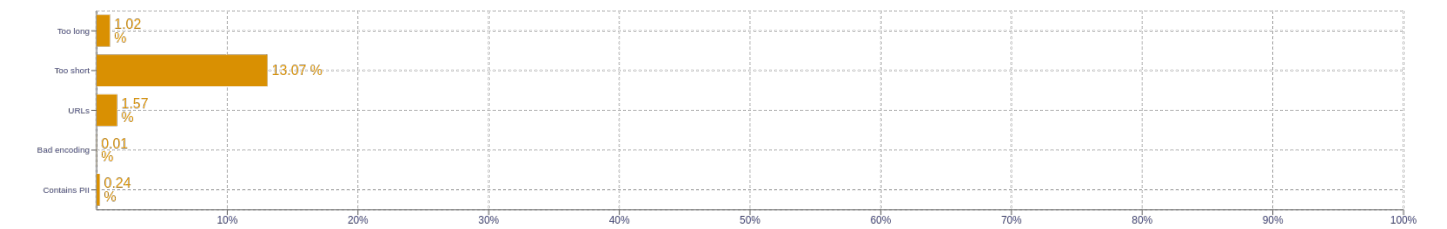
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>word 6406818</div> <div>of 6184511</div> <div>deur 3733665</div> <div>hierdie 3584126</div> <div>ook 3389903</div>
2	<div>of the 474554</div> <div>wysig bron 345356</div> <div>word deur 234204</div> <div>gebruik word 180407</div> <div>moet word 173996</div>
3	<div>vanaf die oorspronklike 48240</div> <div>oor die algemeen 43175</div> <div>sout en peper 36126</div> <div>voor te berei 30939</div> <div>woord van god 30722</div>
4	<div>geargiveer vanaf die oorspronklike 43348</div> <div>wikimedia commons het meer 24986</div> <div>commons het meer media 24983</div> <div>we are searching data 24886</div> <div>searching data for your 24886</div>
5	<div>wikimedia commons het meer media 24983</div> <div>we are searching data for 24886</div> <div>searching data for your request 24886</div> <div>are searching data for your 24886</div> <div>speletjie saam met jou beste 24549</div>

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.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>, , , 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>