HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
zul_Latn.jsonl.tsv	9/19/2024	Zulu (zu)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
113,624	2,710,379	1,461,148 (53.91 %)	56M	362.23 MB	378,209,676

Top 10 domains

Domain	Docs	% of total
news24.com	22K	19.76
voandebele.com	5.4K	4.76
airbnb.com	5.3K	4.69
iol.co.za	4.1K	3.61
jw.org	3.9K	3.39
wordplanet.org	2.9K	2.57
impempe.com	2.2K	1.93
scrolla.africa	2K	1.74
ulwaziprogramme.org	1.5K	1.33
googleplaystoreapks.com	1.3K	1.17

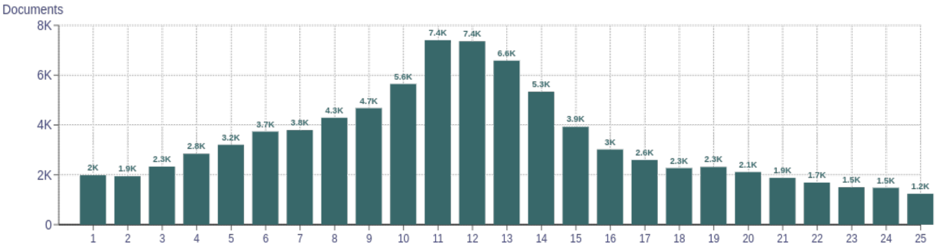
Top 10 TLDs

Domain	Docs	% of total
com	74K	65.26
org	16K	13.67
co.za	10K	9.18
net	2.1K	1.86
africa	2K	1.75
zone	1.1K	0.95
top	804	0.71
info	676	0.59
ac.za	646	0.57
de	588	0.52

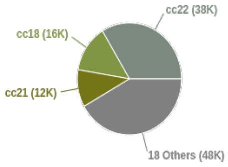
Documents size (in segments)

<= **25** segments **74.94%** (85K documents)

> **25** segments **25.06%** (28K documents)

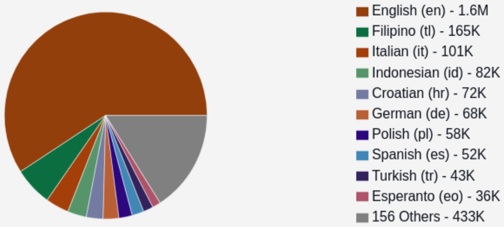


Documents by collection



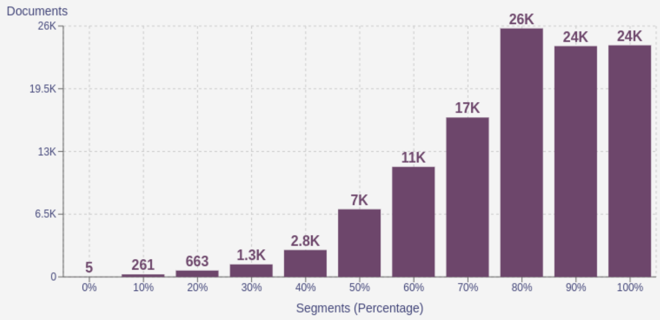
Language Distribution

Number of segments



*Zulu (zu) identification might be inaccurate because language is not supported by Fasttext

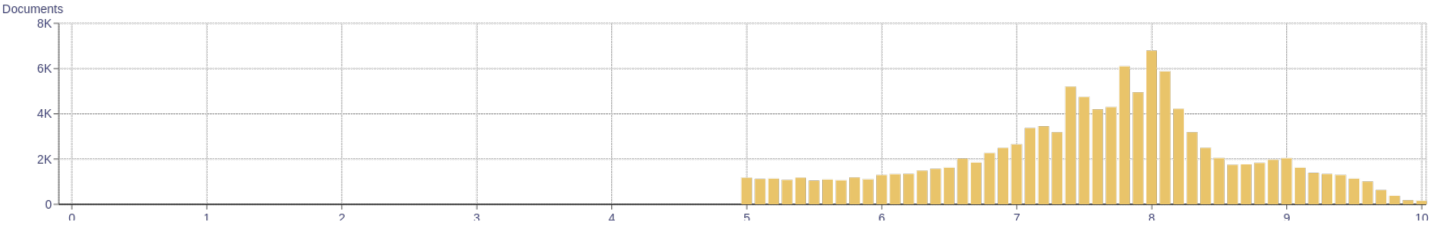
Percentage of segments in Zulu (zu) inside documents



Distribution of documents by document score

score < 5 - **0%** (0 documents)

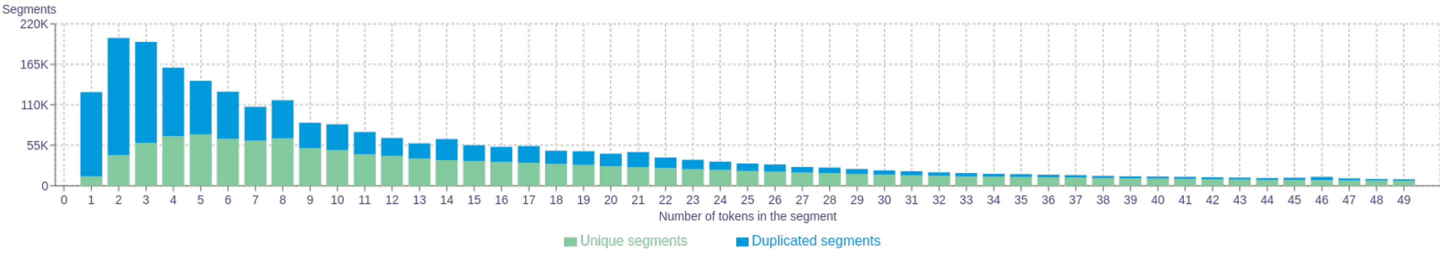
score >= 5 - **100%** (114K documents)



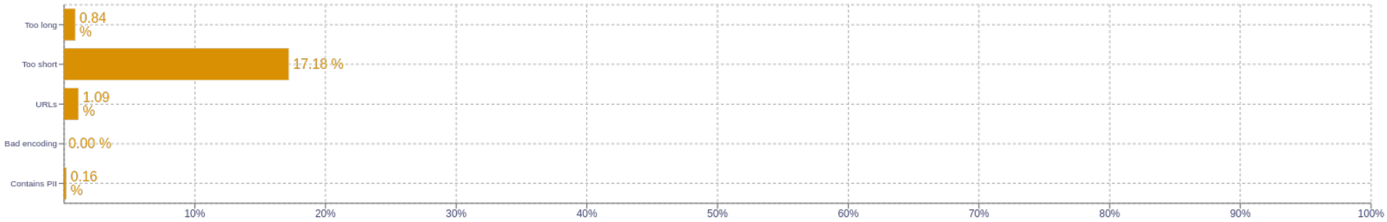
Segment length distribution by token

<= 49 tokens = **1.3M** segments | **1.2M** duplicates

> 50 tokens = **248K** segments | **54K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	i | 393442 noma | 291053 the | 136715 abantu | 128879 lokhu | 106036
2	kuze kube | 15851 ikhasi lethu | 15417 ngolimi lwakho | 15123 of the | 14831 in the | 14807
3	ukwazi ukujoyina inkulumo | 7966 facebook uze ukwazi | 7965 uze ukwazi ukujoyina | 7963 ukujoyina inkulumo ngolimi | 7385 inkulumo ngolimi lwakho | 7258
4	uze ukwazi ukujoyina inkulumo | 7962 facebook uze ukwazi ukujoyina | 7961 ukwazi ukujoyina inkulumo ngolimi | 7381 ukujoyina inkulumo ngolimi lwakho | 7258 kufacebook ukuze uthole izindaba | 7102
5	facebook uze ukwazi ukujoyina inkulumo | 7961 uze ukwazi ukujoyina inkulumo ngolimi | 7377 ukwazi ukujoyina inkulumo ngolimi lwakho | 7254 lethu kufacebook ukuze uthole izindaba | 7101 ikhasi lethu kufacebook ukuze uthole | 7101

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