HPLT Analytics report

General overview

Corpus		Analytics date	Language
nya_Latn.jsonl.tsv		11/11/2024	Nyanja (ny)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
53,117	1,343,861	929,671 (69.18 %)	33M	193.85 MB	201,626,613

Top 10 domains

Domain	Docs	% of total
jw.org	5.8K	10.98
wondershare.com	4.3K	8.07
recetin.com	1.6K	2.94
eturbonews.com	1K	1.94
mwnation.com	994	1.87
martech.zone	930	1.75
androidsis.com	903	1.70
heartsandlaserbeams.com	792	1.49
wikipedia.org	607	1.14
biblica.com	541	1.02

Top 10 TLDs

Domain	Docs	% of total
com	39K	73.98
org	8.3K	15.64
zone	930	1.75
net	806	1.52
ru	340	0.64
xyz	326	0.61
mw	245	0.46
es	231	0.43
pt	215	0.40
news	181	0.34

Documents size (in segments)

<= 25 segments **71.87%** (38K documents)
> 25 segments **28.13%** (15K documents)
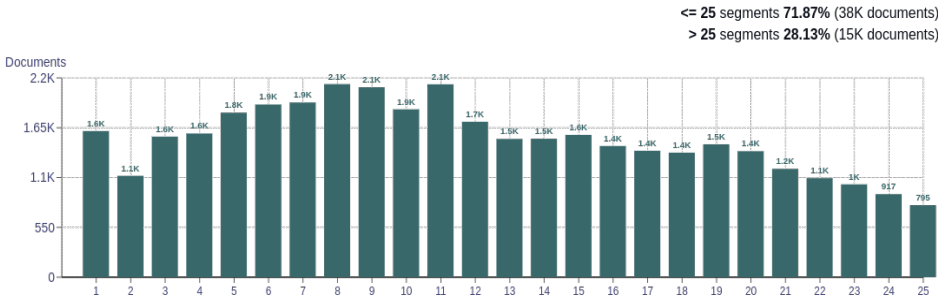
Documents by collection

cc22 (27K)
cc21 (8K)
19 Others (19K)
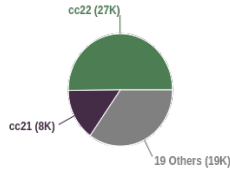
Language Distribution

Number of segments

- English (en) - 626K
- Swahili (sw) - 137K
- Filipino (tl) - 78K
- Croatian (hr) - 66K
- Italian (it) - 64K
- Polish (pl) - 57K
- Spanish (es) - 53K
- Indonesian (id) - 53K
- Esperanto (eo) - 31K
- German (de) - 24K
- 151 Others - 156K

*Nyanja (ny) identification might be inaccurate because language is not supported by Fasttext
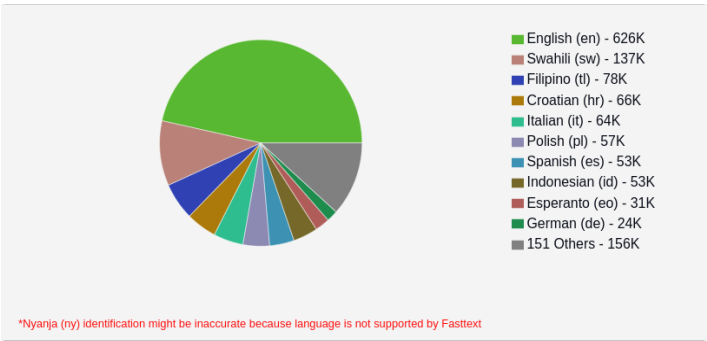
### Percentage of segments in Nyanja (ny) inside documents
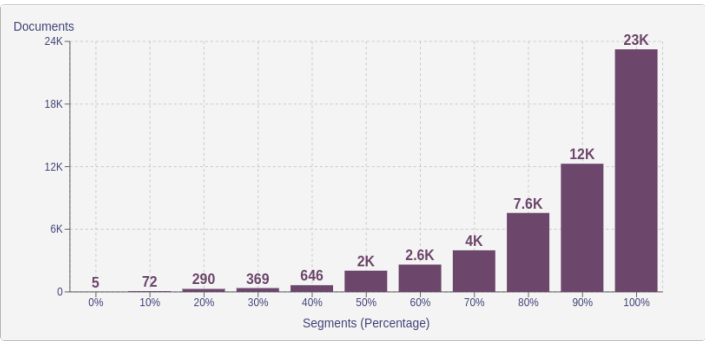
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (53K documents)

Segment length distribution by token

<= 49 tokens = **769K** segments | **383K** duplicates
> 50 tokens = **192K** segments | **31K** duplicates

- Unique segments
- Duplicated segments

Segment noise distribution

Too long	0.00 %
Too short	11.20 %
URLs	1.17 %
Bad encoding	0.00 %
Contains PII	0.12 %
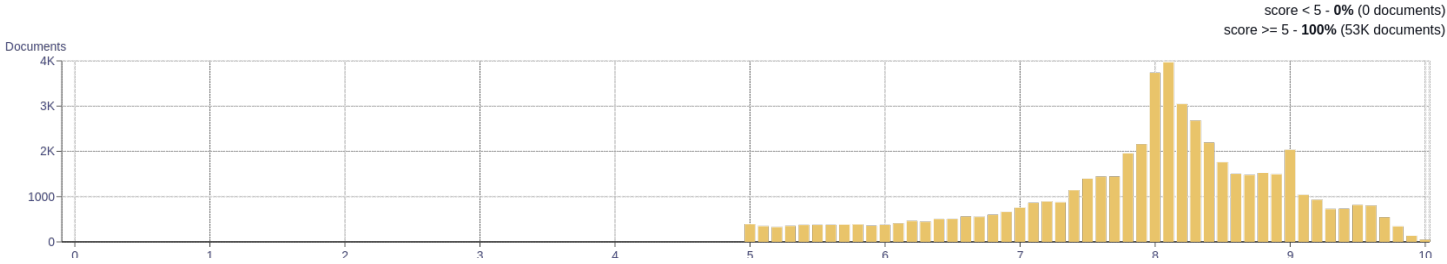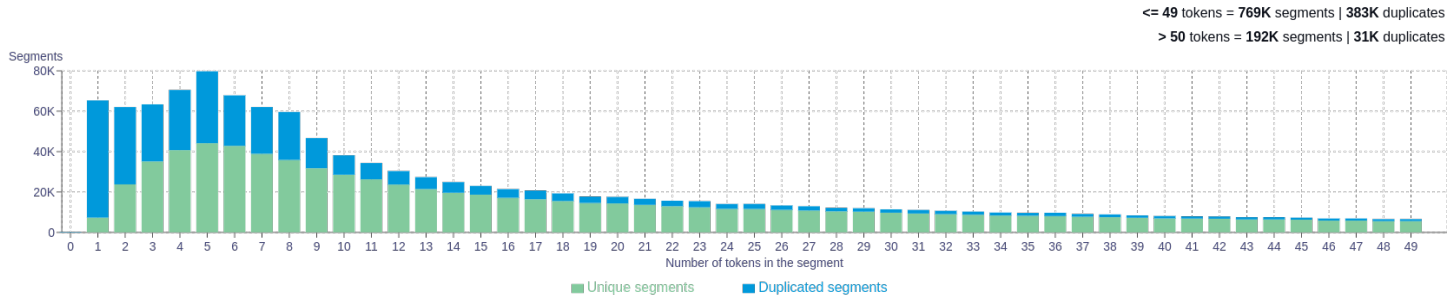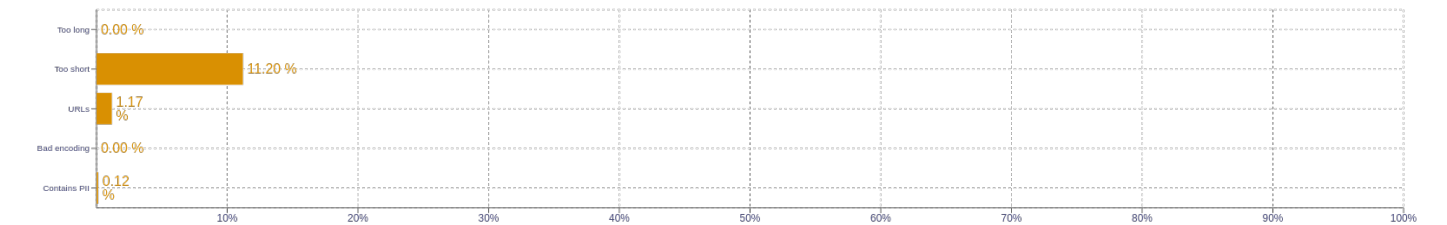
About HPLT Analytics

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