HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
lug_Latn.jsonl.tsv	9/19/2024	Ganda (lg)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
21,276	407,541	240,013 (58.89 %)	12M	65.91 MB	67,578,510

Top 10 domains

Domain	Docs	% of total
bukedde.co.ug	4.2K	19.81
bible.is	2.2K	10.37
dembefm.ug	1.9K	9.01
wordplanet.org	1.9K	8.72
wordproject.org	1.5K	7.13
jw.org	1.4K	6.71
radiosimba.ug	1.1K	5.17
cbsfm.ug	973	4.57
nbs.ug	576	2.71
wikipedia.org	394	1.85

Top 10 TLDs

Domain	Docs	% of total
org	5.6K	26.53
co.ug	5.6K	26.34
ug	4.7K	22.05
com	2.7K	12.75
is	2.2K	10.37
net	165	0.78
info	36	0.17
eu	31	0.15
or.ug	22	0.10
ca	15	0.07

Documents size (in segments)

<= 25 segments **80.11%** (17K documents)
> 25 segments **19.89%** (4.2K documents)
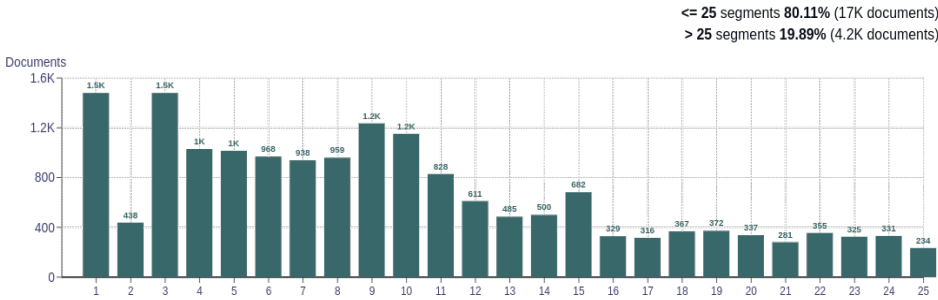
Documents by collection

cc18 (2.4K) cc22 (5.8K)
19 Others (13K)
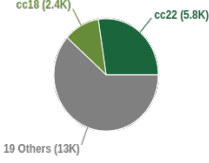
Language Distribution

Number of segments

- English (en) - 128K
- Waray (war) - 28K
- Finnish (fi) - 26K
- Swahili (sw) - 25K
- Filipino (tl) - 23K
- Esperanto (eo) - 21K
- Sundanese (su) - 18K
- Indonesian (id) - 13K
- German (de) - 13K
- Spanish (es) - 11K
- 138 Others - 101K

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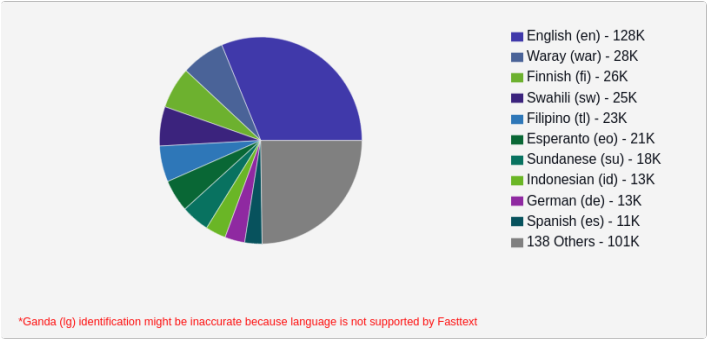
Percentage of segments in Ganda (lg) inside documents
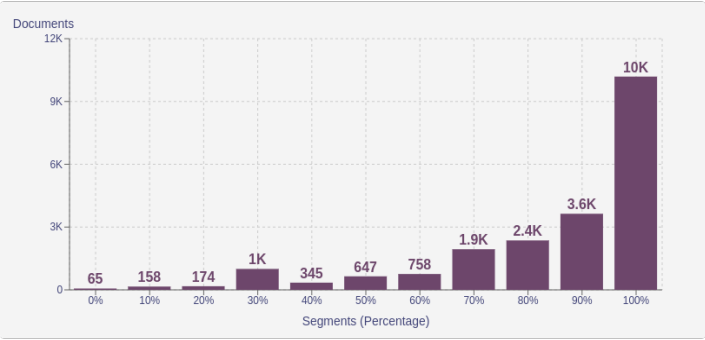
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (21K documents)

Segment length distribution by token

<= 49 tokens = **207K** segments | **153K** duplicates
> 50 tokens = **48K** segments | **15K** duplicates

- Unique segments
- Duplicated segments

Segment noise distribution

Too long	1.11 %
Too short	10.44 %
URLs	0.39 %
Bad encoding	0.01 %
Contains PII	0.13 %
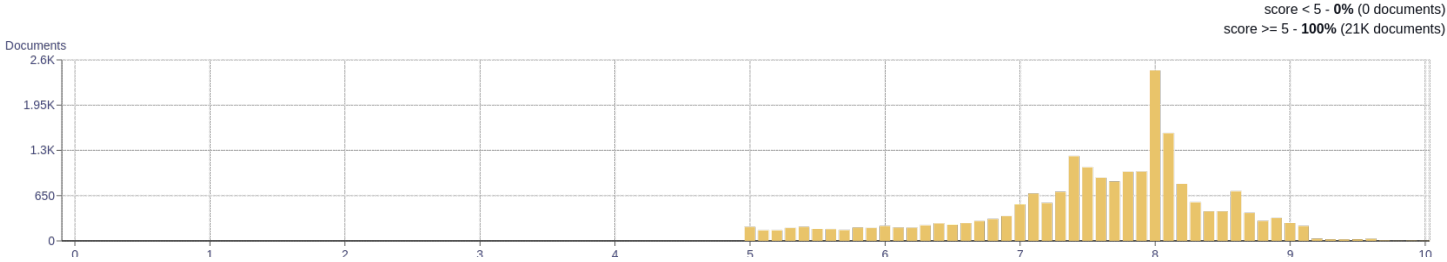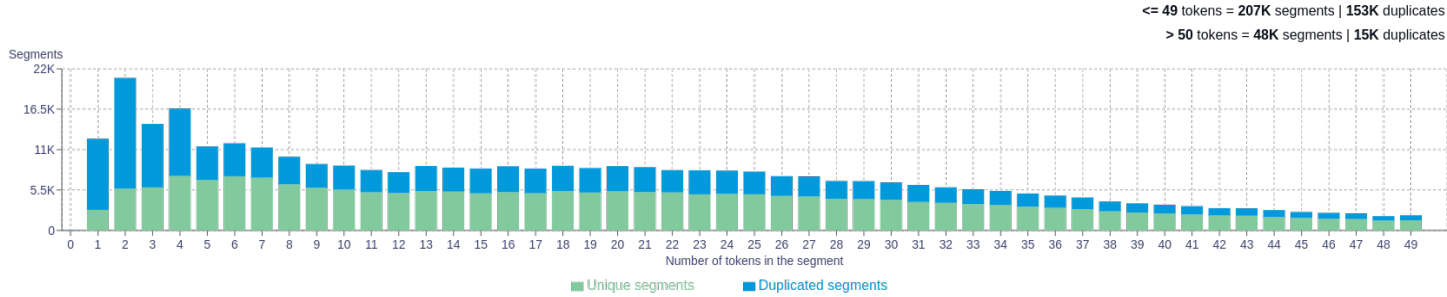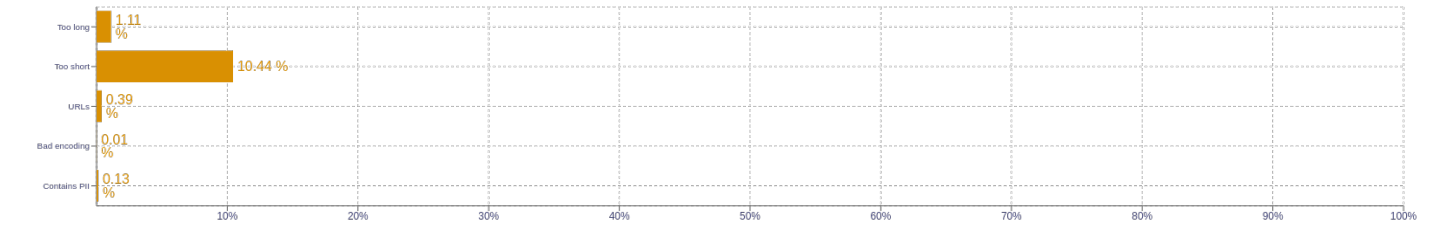
Frequent n-grams

Size	n-grams
1	n | 268953 ng | 57960 omu | 50175 abantu | 38919 bw | 36845
2	mukama katonda | 4401 wamu n | 3041 yesu kristo | 2983 mukama n | 2680 mukama waffe | 2457
3	abaana ba isiraeri | 1757 maaso ga mukama | 1352 mukama katonda wo | 1065 ekigambo kya mukama | 823 mukama waffe yesu | 679
4	bye bye bye bye | 520 empapula eziriko enyunzi ezigguka | 449 mukama waffe yesu kristo | 434 enyunzi ezigguka ku luno | 391 endagaano enkadde n'endagaano empya | 389
5	bye bye bye bye bye | 490 eziriko enyunzi ezigguka ku luno | 391 do you have a story | 254 your community or an opinion | 253 you have a story in | 253

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