HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
gaz_Latn.jsonl.tsv	12/16/2024	Oromo (gaz)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
49,139	973,633	525,017 (53.92 %)	36M	212.25 MB	218,282,982

Top 10 domains

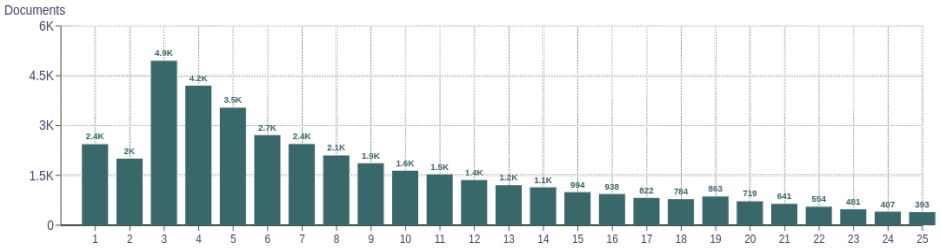
Domain	Docs	% of total
voaafaanoromoo.com	11K	22.39
bible.is	3.3K	6.62
qeerroo.org	2.7K	5.56
bilisummaa.com	2.7K	5.51
nuuralhudaa.com	1.9K	3.85
gadaa.com	1.7K	3.47
wikipedia.org	1.6K	3.18
sammubani.com	1.5K	2.95
kichuu.com	1.4K	2.79
ayyaantuu.org	1.4K	2.76

Top 10 TLDs

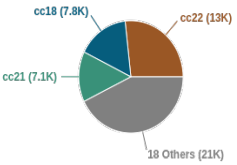
Domain	Docs	% of total
com	30K	61.97
org	10K	21.29
is	3.3K	6.62
no	1.1K	2.34
net	713	1.45
et	593	1.21
gov.et	388	0.79
de	342	0.70
dk	290	0.59
gov	218	0.44

Documents size (in segments)

<= 25 segments **82.84%** (41K documents)
> 25 segments **17.16%** (8.4K documents)

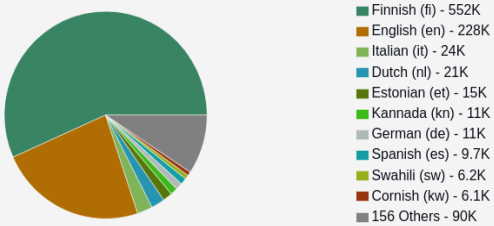


Documents by collection



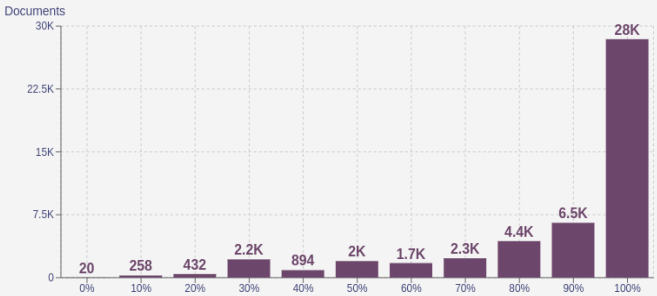
Language Distribution

Number of segments



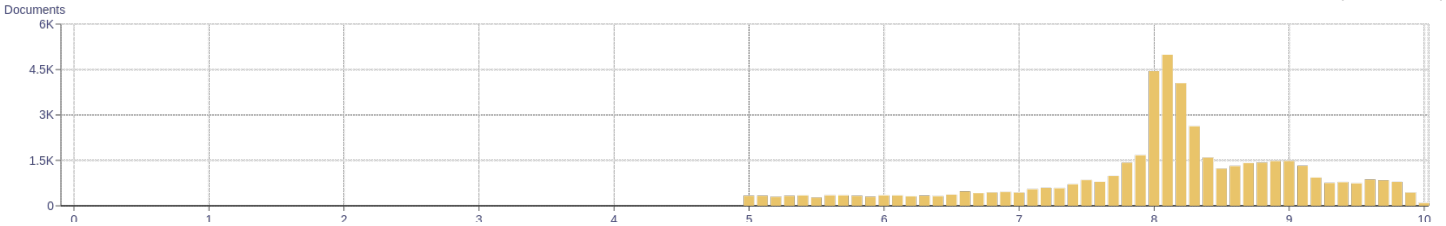
*Oromo (gaz) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Oromo (gaz) inside documents



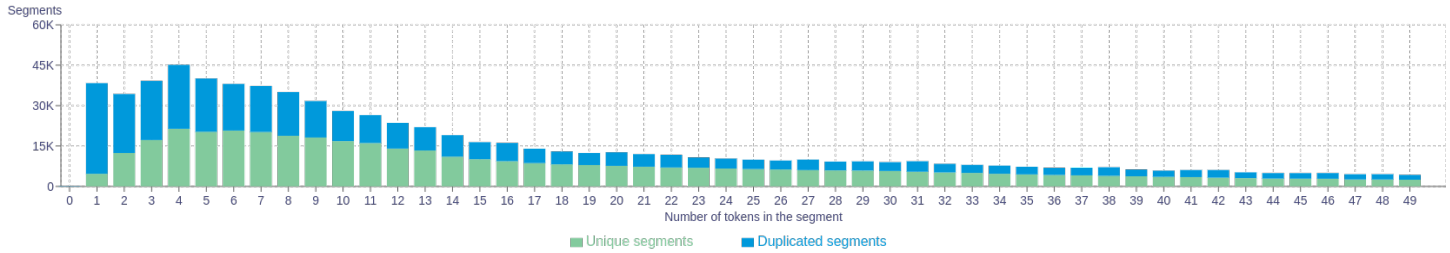
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (49K documents)

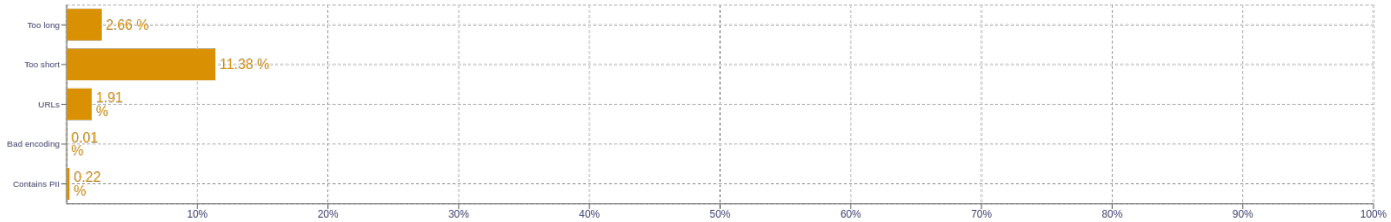


Segment length distribution by token

<= 49 tokens = **412K** segments | **353K** duplicates
> 50 tokens = **209K** segments | **96K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ta | 266637 · oromoo | 206247 · irratti | 136717 · keessatti | 135879 · irraa | 134479
2	of the | 17118 · bilisummaa oromoo | 16589 · adda addaa | 14969 · ummata oromoo | 13337 · ilmaan oromoo | 12954
3	osoo hin taane | 5083 · baaa baaa baaa | 2844 · subhaanahu wa ta | 2696 · qeerroo bilisummaa oromoo | 2629 · adda bilisummaa oromoo | 2587
4	baaa baaa baaa baaa | 2841 · rabbiin subhaanahu wa ta | 2165 · qofa osoo hin taane | 1372 · reserves the right to | 583 · do not follow the | 580
5	baaa baaa baaa baaa baaa | 2838 · you take all the responsibility | 578 · unsubstantiated allegations are not allowed | 578 · to remove comments that do | 578 · the right to remove comments | 578

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