HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
war_Latn.jsonl.tsv	10/31/2024	Waray (war)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
13,873	200,935	87,226 (43.41 %)	7.2M	33.95 MB	35,387,743

Top 10 domains

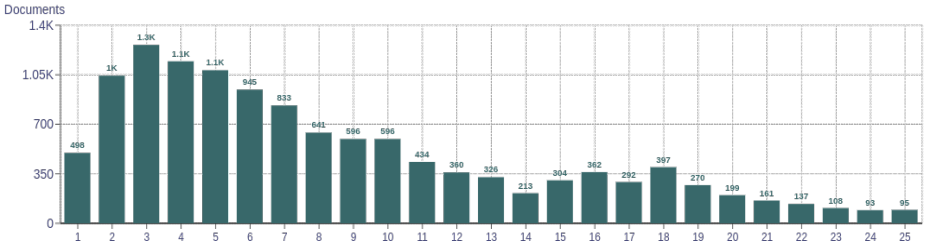
Domain	Docs	% of total
wikipedia.org	10K	74.15
bible.is	735	5.30
jw.org	537	3.87
isumat.com	410	2.96
info-about.ru	324	2.34
bomboradyo.com	291	2.10
pia.gov.ph	169	1.22
rmn.ph	122	0.88
tacloban.gov.ph	112	0.81
wordpress.com	89	0.64

Top 10 TLDs

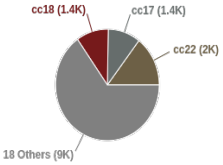
Domain	Docs	% of total
org	11K	79.08
com	1.2K	8.53
is	735	5.30
gov.ph	340	2.45
ru	326	2.35
ph	136	0.98
net	34	0.25
click	26	0.19
de	22	0.16
info	11	0.08

Documents size (in segments)

<= 25 segments **89.32%** (12K documents)
> 25 segments **10.68%** (1.5K documents)

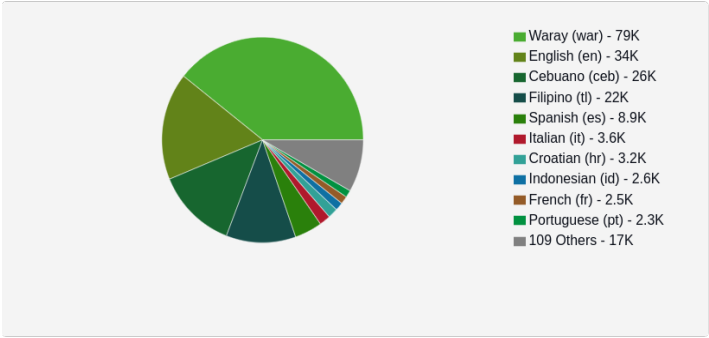


Documents by collection

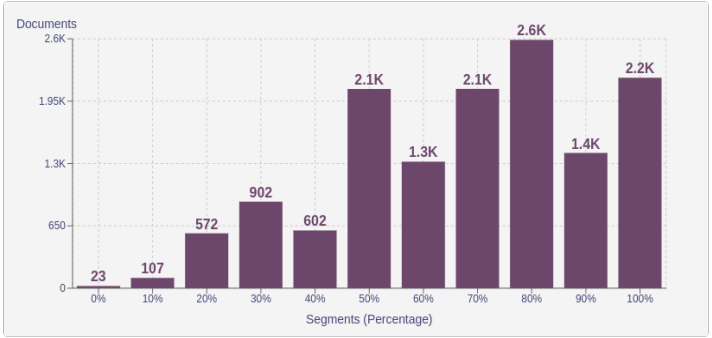


Language Distribution

Number of segments

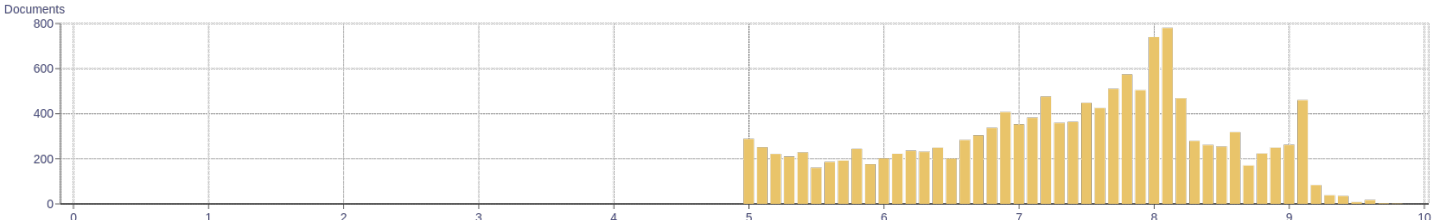


Percentage of segments in Waray (war) inside documents



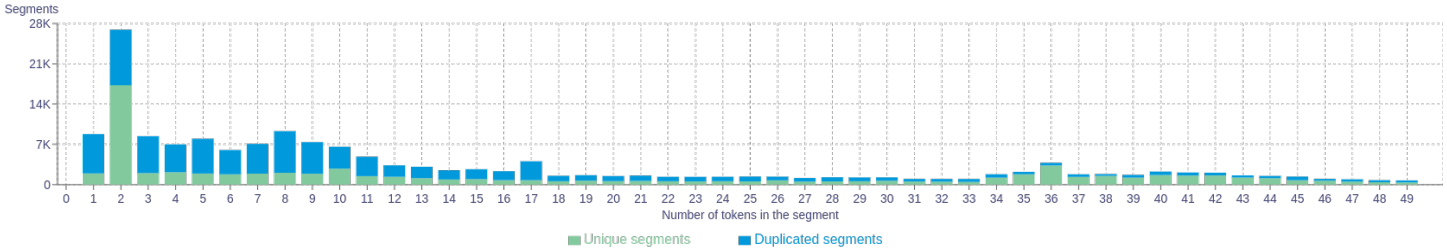
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (14K documents)

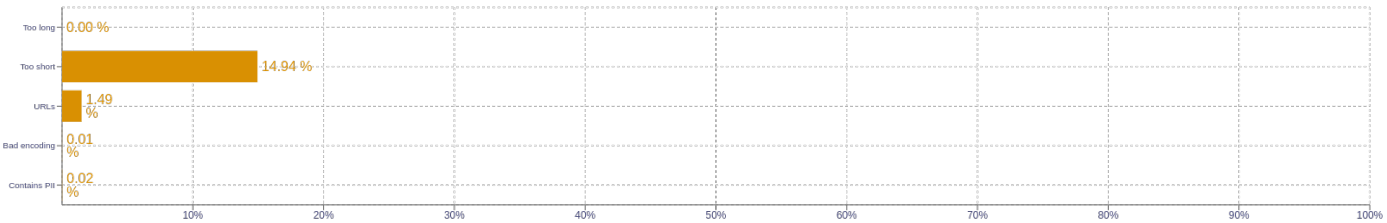


Segment length distribution by token

<= 49 tokens = **73K** segments | **94K** duplicates
> 50 tokens = **34K** segments | **20K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ha | 552950 · nga | 434930 · han | 309581 · an | 263731 · mga | 155874
2	ha ha | 354612 · nga mga | 37392 · han mga | 26744 · an mga | 19438 · amo an | 16189
3	ha ha ha | 353243 · in nahilalakip ha | 15635 · uska species han | 14402 · in uska species | 14396 · nahilalakip ha genus | 13405
4	ha ha ha ha | 351920 · in uska species han | 14396 · in nahilalakip ha genus | 13405 · nahilalakip ha genus nga | 13285 · uska species han magnoliopsida | 9694
5	ha ha ha ha ha | 350666 · in nahilalakip ha genus nga | 13285 · in uska species han magnoliopsida | 9694 · uska species han magnoliopsida nga | 9630 · species han magnoliopsida nga ginhulagway | 7486

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