HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
pag_Latn.jsonl.tsv	12/3/2024	Pangasinan (pag)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
6,900	85,832	56,530 (65.86 %)	6.9M	32.28 MB	33,444,943

Top 10 domains

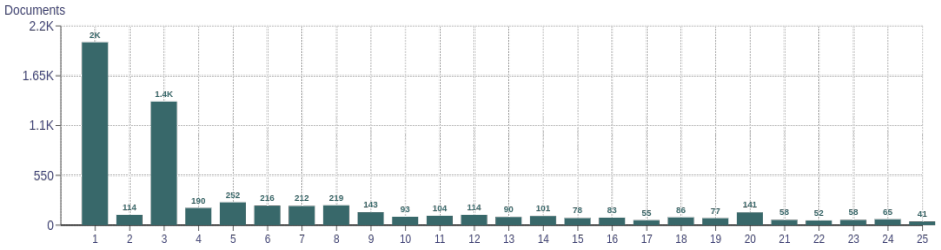
Domain	Docs	% of total
bible.is	3K	43.65
jw.org	1.9K	27.33
wikipedia.org	424	6.14
bomboradyo.com	338	4.90
blogspot.jp	137	1.99
ebible.org	120	1.74
blogspot.com	84	1.22
wordpress.com	76	1.10
online.pk	60	0.87
tuugo.ph	50	0.72

Top 10 TLDs

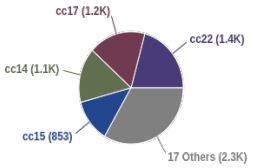
Domain	Docs	% of total
is	3K	43.65
org	2.5K	36.65
com	912	13.22
jp	138	2.00
pk	60	0.87
ph	51	0.74
net	37	0.54
in	20	0.29
eu	14	0.20
co.uk	11	0.16

Documents size (in segments)

<= 25 segments **87.33%** (6K documents)
> 25 segments **12.67%** (874 documents)

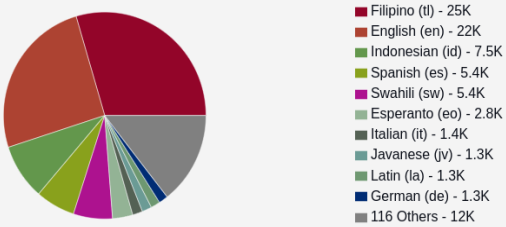


Documents by collection



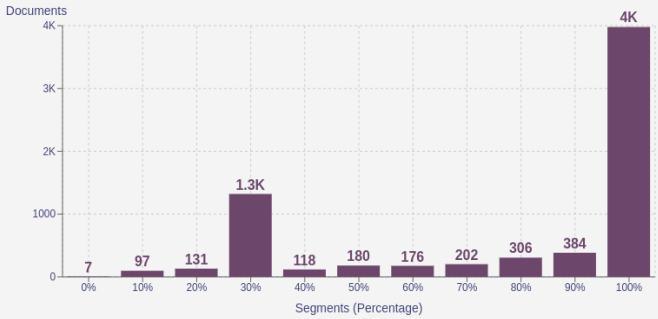
Language Distribution

Number of segments



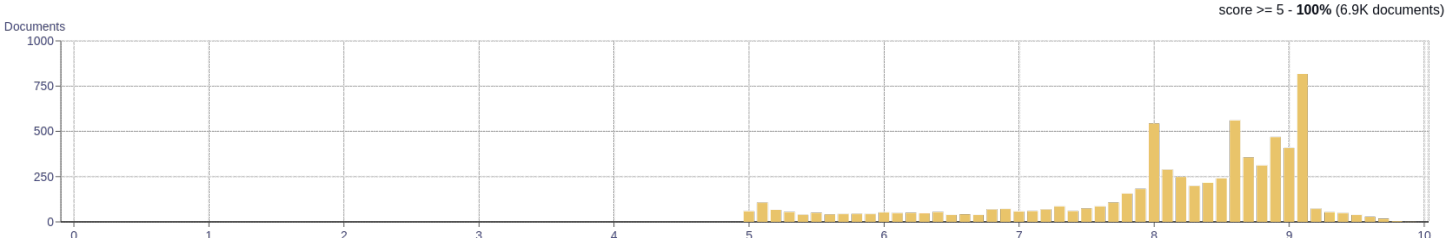
*Pangasinan (pag) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Pangasinan (pag) inside documents



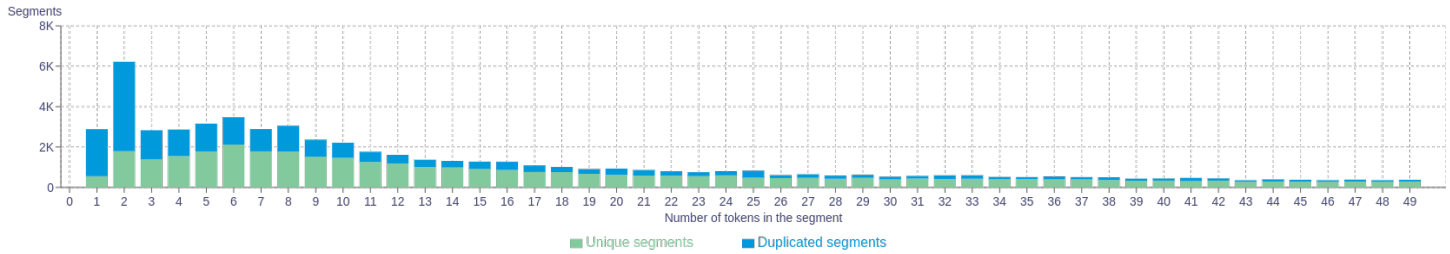
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (6.9K documents)

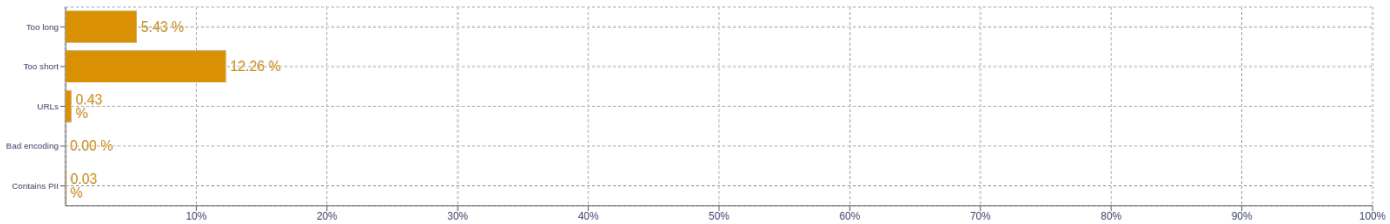


Segment length distribution by token

<= 49 tokens = **36K** segments | **24K** duplicates
> 50 tokens = **26K** segments | **5.5K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ya | 239660 · ed | 104934 · so | 101201 · tan | 68452 · et | 56673
2	ed saray | 10221 · nen jehova | 7788 · nen jesus | 7469 · nin diyos | 6636 · apu dios | 5687
3	y afu dios | 3211 · iran iran iran | 2102 · ed si jehova | 2040 · hi apo hisos | 1762 · ibabaga na biblia | 1208
4	iran iran iran iran | 2094 · saray tasi nen jehova | 668 · ya innia laman nga | 566 · manga oripn o allāh | 530 · y afu dios nga | 504
5	iran iran iran iran iran | 2086 · so so so so so | 399 · pinili ni apo namalyari hên | 286 · mánoron kautuhan ni apo moises | 196 · ya panakallan ira kan ya | 185

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