HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-v2-ind_Latn.tsv	10/4/2024	Indonesian (id)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
98,141,856	2,388,663,752			357.31 GB	381,930,808,361

Top 10 domains

Domain	Docs	% of total
blogspot.com	8M	8.17
wordpress.com	3.7M	3.77
tribunnews.com	1.6M	1.68
blogspot.co.id	1.3M	1.31
kompas.com	783K	0.80
blogspot.sg	470K	0.48
us.com	406K	0.41
detik.com	404K	0.41
wikipedia.org	359K	0.37
okezone.com	332K	0.34

Top 10 TLDs

Domain	Docs	% of total
com	66M	66.77
co.id	6M	6.07
net	4M	4.09
id	3.9M	4.02
org	3.2M	3.24
info	1.8M	1.85
co	1.7M	1.72
ac.id	1.6M	1.68
go.id	1.5M	1.51
web.id	1.1M	1.07

Documents size (in segments)

<= 25 segments **76.95%** (76M documents)
> 25 segments **23.05%** (23M documents)
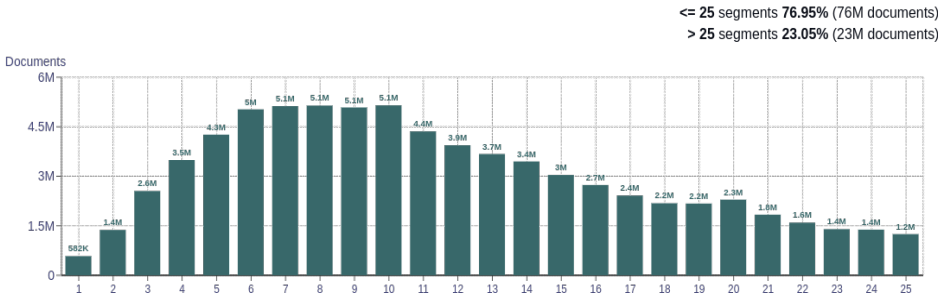
Documents by collection

cc18 (19M)
cc22 (24M)
cc21 (12M)
18 Others (43M)
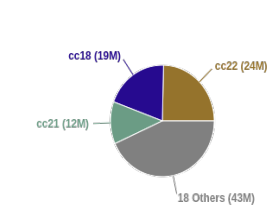
Language Distribution

Number of segments

- Indonesian (id) - 1.9B
- English (en) - 224M
- Malay (ms) - 76M
- Italian (it) - 58M
- French (fr) - 24M
- Spanish (es) - 15M
- German (de) - 15M
- Dutch (nl) - 12M
- Swedish (sv) - 9.8M
- Esperanto (eo) - 6.8M
- 165 Others - 88M
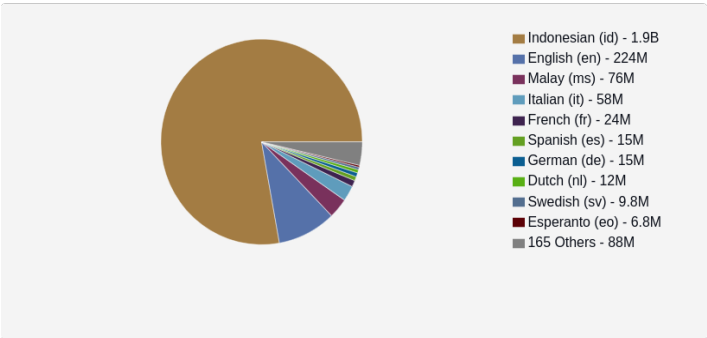
Percentage of segments in Indonesian (id) inside documents
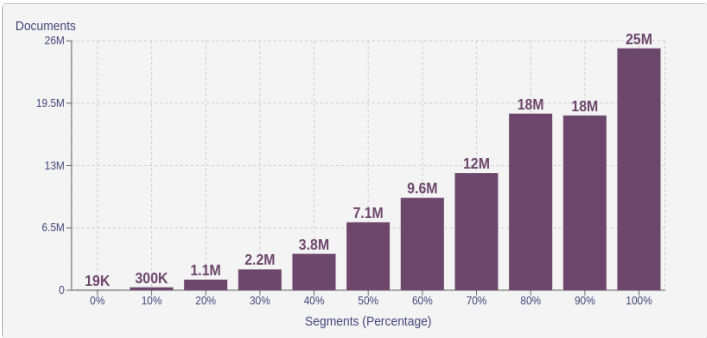
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (98M documents)

Segment noise distribution

Too long	0.00 %
Too short	5.28 %
URLs	2.67 %
Bad encoding	0.01 %
Contains PII	0.15 %
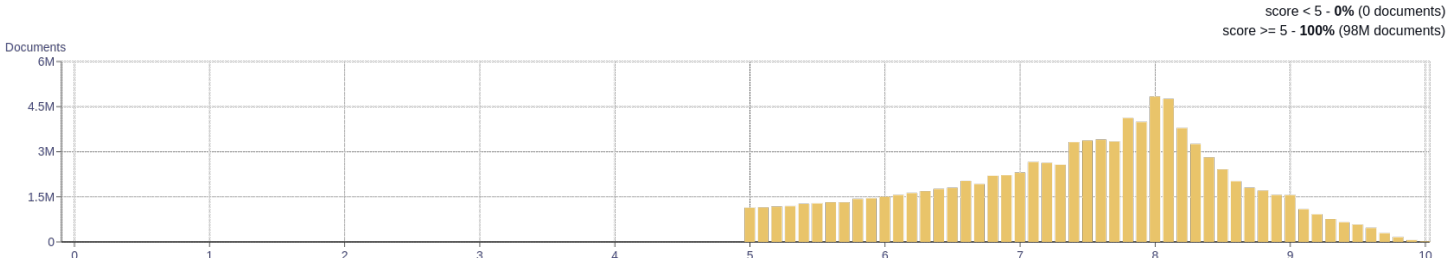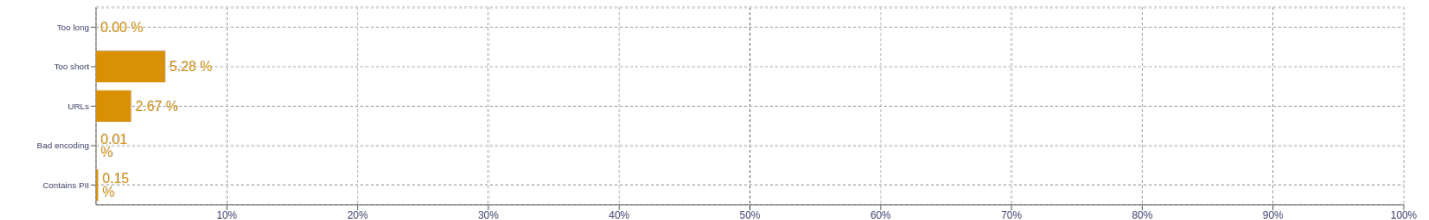
About HPLT Analytics

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