HPLT Analytics report

General overview

Corpus	Analytics date	Language
HPLT-docslite.id.tsv	6/28/2024	Indonesian (id)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
31,420,597	4,129,834,951			270.0 GB	

Top 10 domains

Domain	Docs	% of total
blogspot.co.id	1.7M	5.29
blogspot.com	1.3M	4.27
blogspot.sg	582K	1.85
us.com	379K	1.21
alibaba.com	379K	1.20
petalokasi.org	252K	0.80
blogspot.kr	189K	0.60
wordpress.com	182K	0.58
aliexpress.com	180K	0.57
blogspot.jp	161K	0.51

Top 10 TLDs

Domain	Docs	% of total
com	18M	58.09
co.id	2.6M	8.16
net	1.2M	3.83
org	1.2M	3.68
id	1.1M	3.35
sg	583K	1.85
club	576K	1.83
info	560K	1.78
co	554K	1.76
ac.id	421K	1.34

Documents size (in segments)

<= 25 segments **7.77%** (2.4M documents)
> 25 segments **92.23%** (29M documents)
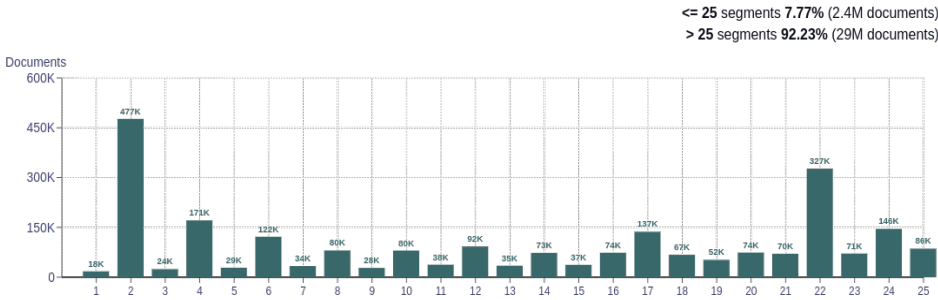
Documents by collection

cc40 (10M), wide15 (8.4M), wide17 (5.3M), wide16 (7.3M)
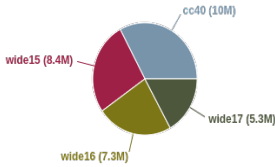
Language Distribution

Number of segments

- Indonesian (id) - 1.7B
- English (en) - 1.3B
- Malay (ms) - 220M
- German (de) - 88M
- Italian (it) - 85M
- Czech (cs) - 84M
- French (fr) - 80M
- Spanish (es) - 67M
- Swedish (sv) - 52M
- Dutch (nl) - 50M
- 164 Others - 358M
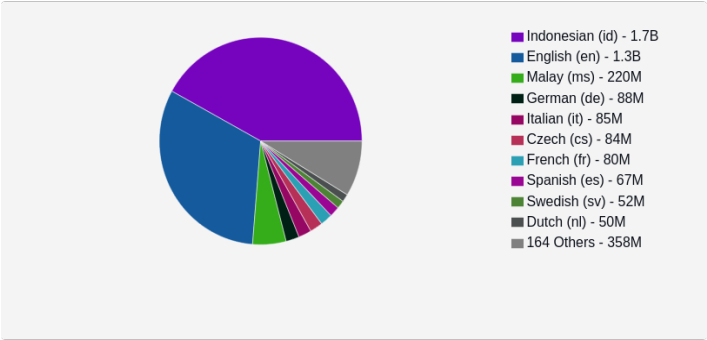
Percentage of segments in Indonesian (id) inside documents
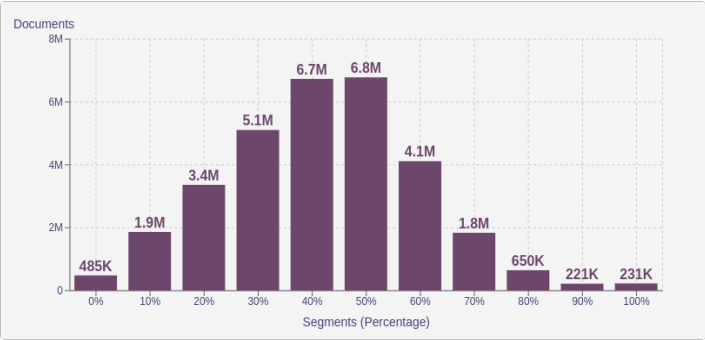
Distribution of documents by document score

score < 5 - **35.22%** (11M documents)
score >= 5 - **64.78%** (20M documents)

Segment noise distribution

- Too long: 0.26 %
- Too short: 33.71 %
- URLs: 3.01 %
- Bad encoding: 0.00 %
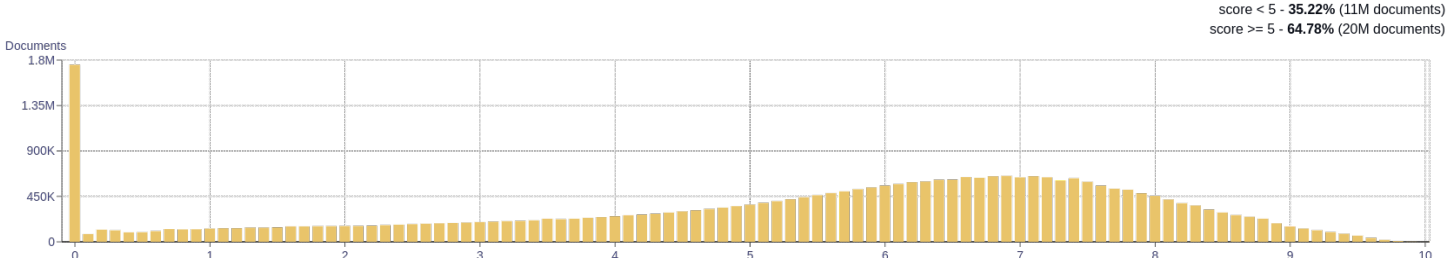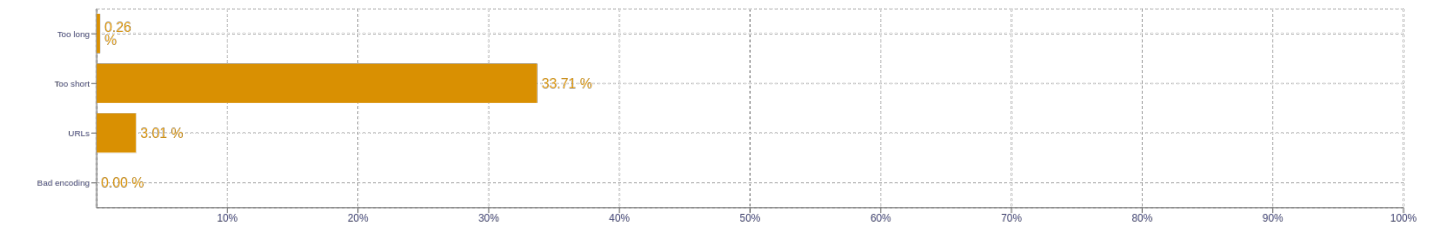
About HPLT Analytics

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