HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
bug_Latn.jsonl.tsv	12/3/2024	Buginese (bug)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
2,023	38,551	24,329 (63.11 %)	3.9M	18.7 MB	19,276,303

Top 10 domains

Domain	Docs	% of total
bible.is	657	32.48
alkitab.mobi	357	17.65
blogspot.com	111	5.49
wikipedia.org	77	3.81
teluguserialonline.com	58	2.87
petalokasi.org	52	2.57
alkitab.pw	49	2.42
wordpress.com	25	1.24
scribd.com	23	1.14
blogspot.co.id	20	0.99

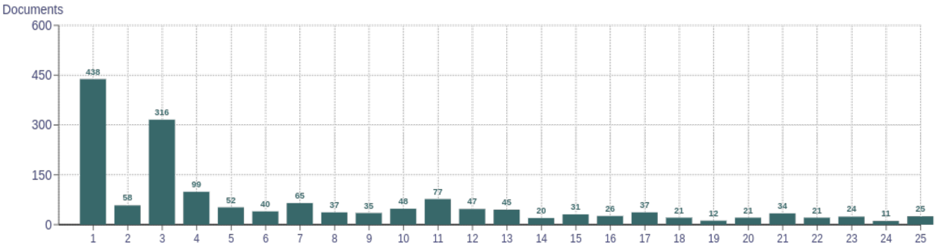
Top 10 TLDs

Domain	Docs	% of total
is	657	32.48
com	551	27.24
mobi	357	17.65
org	188	9.29
net	57	2.82
pw	49	2.42
co.id	25	1.24
info	13	0.64
id	13	0.64
tv	11	0.54

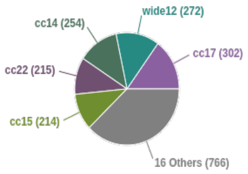
Documents size (in segments)

<= 25 segments **81.07%** (1.6K documents)

> 25 segments **18.93%** (383 documents)

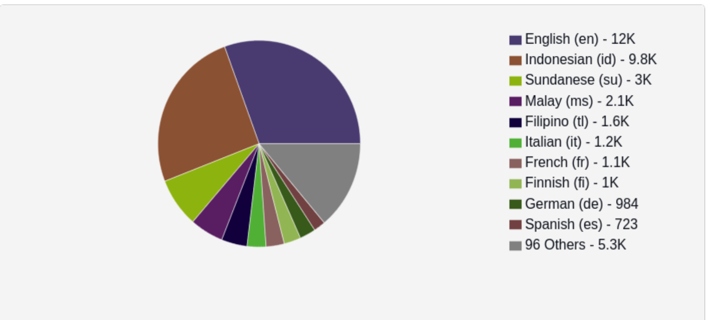


Documents by collection



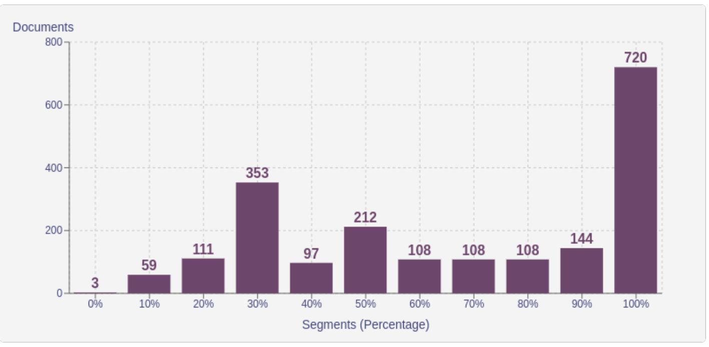
Language Distribution

Number of segments



*Buginese (bug) identification might be inaccurate because language is not supported by Fasttext

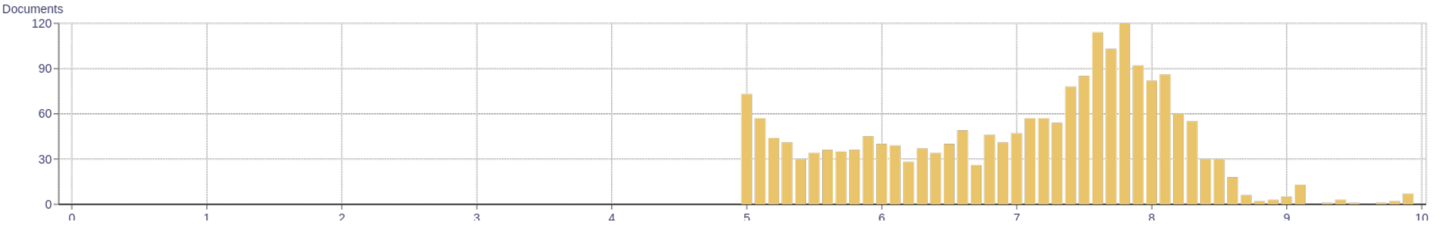
Percentage of segments in Buginese (bug) inside documents



Distribution of documents by document score

score < 5 - **0%** (0 documents)

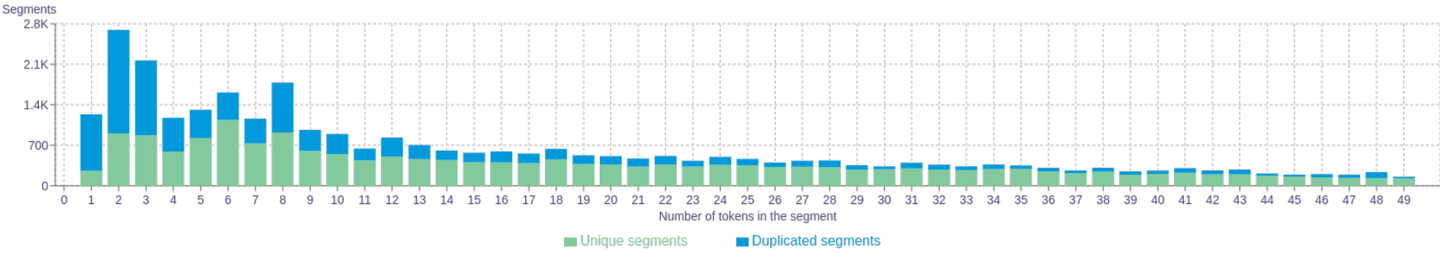
score >= 5 - **100%** (2K documents)



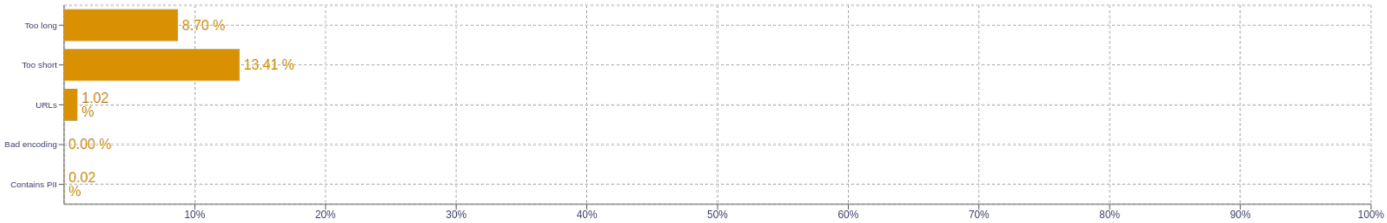
Segment length distribution by token

<= 49 tokens = **19K** segments | **12K** duplicates

> 50 tokens = **7.8K** segments | **2.4K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	i | 40616 anna | 22670 lako | 16997 allataala | 15988 yesus | 15721
2	masser masser | 5719 i yesus | 5497 nu alatala | 4350 kodala kodala | 1850 kodala koduku | 1828
3	masser masser masser | 5718 kodala kodala koduku | 1820 kodala koduku pellama | 1719 wedding wedding wedding | 1524 sampe ri kasae | 648
4	masser masser masser masser | 5717 kodala kodala koduku pellama | 1711 wedding wedding wedding wedding | 1518 episode kodala kodala koduku | 475 nosa gasa nu alatala | 263
5	masser masser masser masser masser | 5716 wedding wedding wedding wedding wedding | 1512 episode kodala kodala koduku pellama | 475 kodala kodala koduku pellama episode | 261 lomba guru memperingati hut pgri | 240

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