

General overview

Corpus	Analytics date	Language
bug_Latn.jsonl.tsv	10/31/2024	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

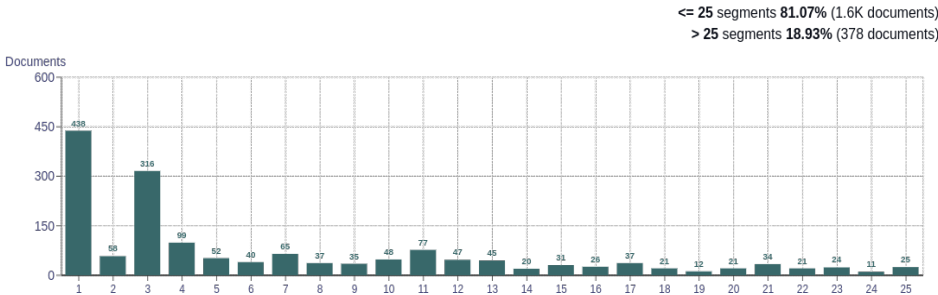
Top 10 domains

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

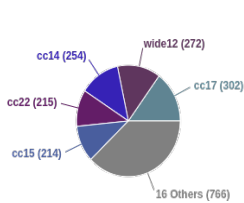
Top 10 TLDs

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)

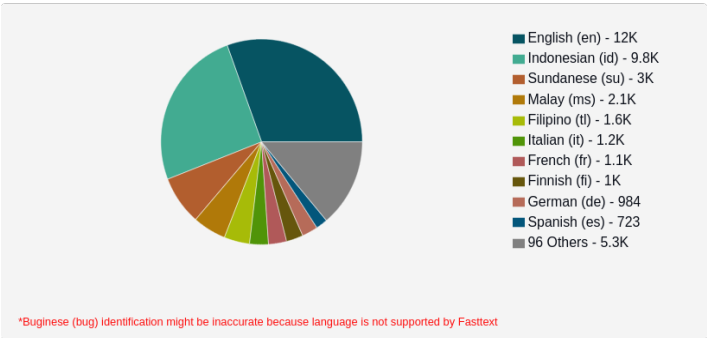


Documents by collection

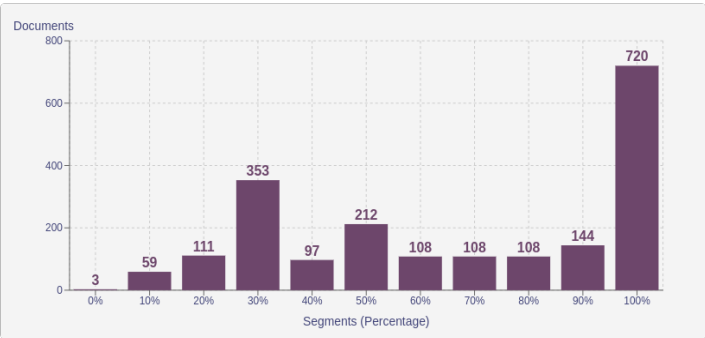


Language Distribution

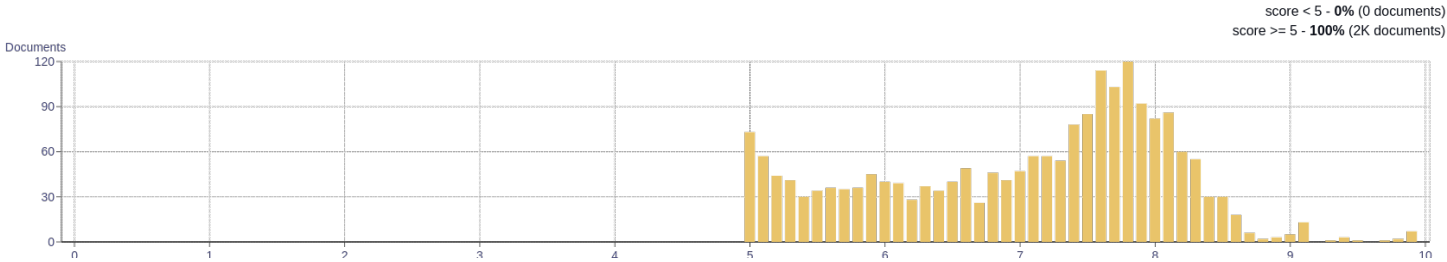
Number of segments



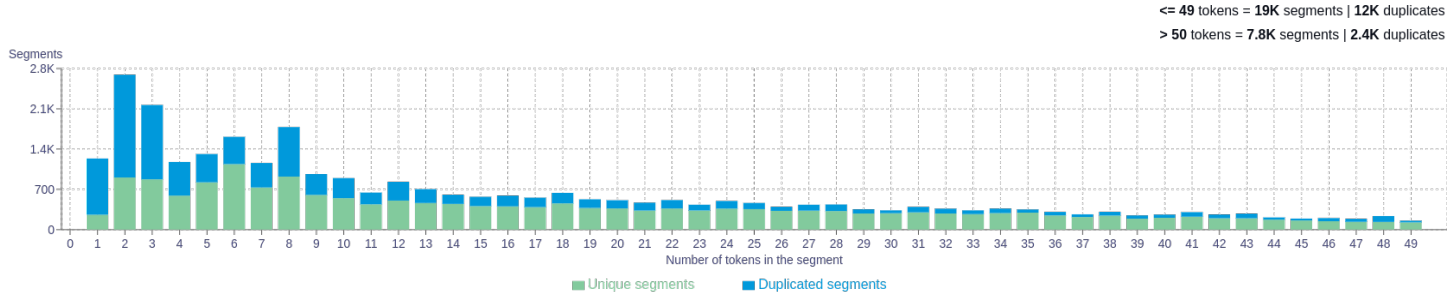
Percentage of segments in Buginese (bug) inside documents



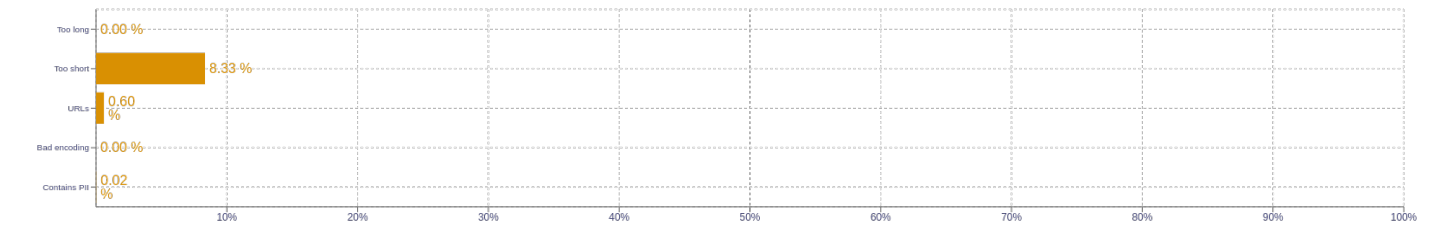
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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>