HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
ban_Latn.jsonl.tsv	10/3/2024	Balinese (ban)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
10,700	601,139	161,172 (26.81 %)	14M	73.75 MB	76,653,907

Top 10 domains

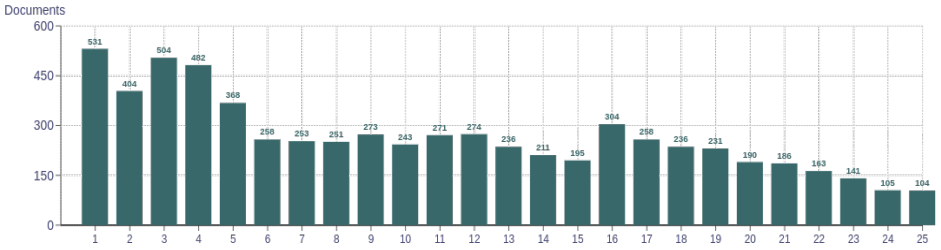
Domain	Docs	% of total
basabali.org	3.1K	28.95
wikipedia.org	1.6K	14.59
blogspot.com	1.5K	13.56
bible.is	674	6.30
wordpress.com	450	4.21
blogspot.co.id	409	3.82
alkitab.mobi	220	2.06
suarasakingbali.com	143	1.34
belajarbahasabali.com	123	1.15
scribd.com	90	0.84

Top 10 TLDs

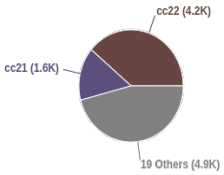
Domain	Docs	% of total
org	5K	46.33
com	3.6K	33.89
is	674	6.30
co.id	429	4.01
mobi	220	2.06
net	166	1.55
ac.id	56	0.52
asia	36	0.34
in	33	0.31
nl	31	0.29

Documents size (in segments)

<= 25 segments **62.36%** (6.7K documents)
> 25 segments **37.64%** (4K documents)

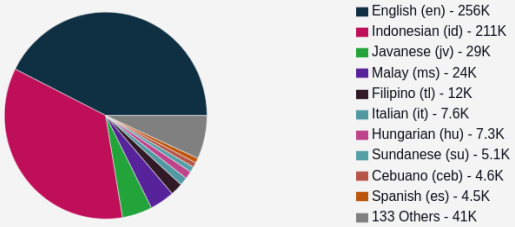


Documents by collection



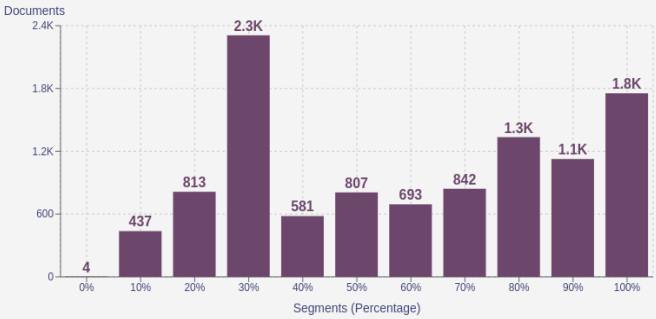
Language Distribution

Number of segments



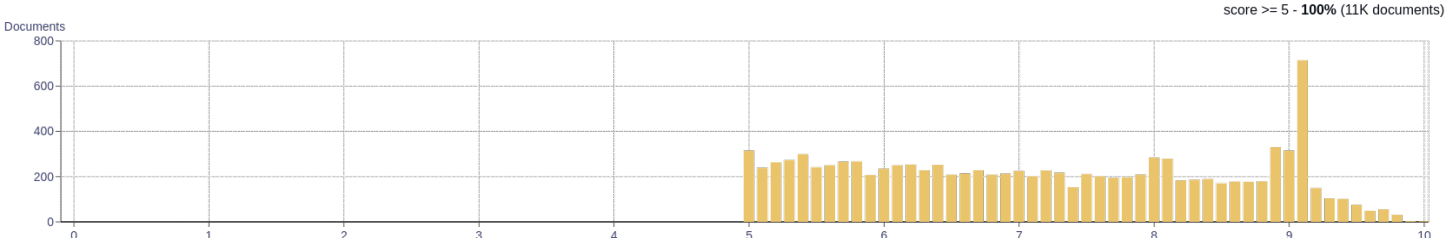
*Balinese (ban) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Balinese (ban) inside documents



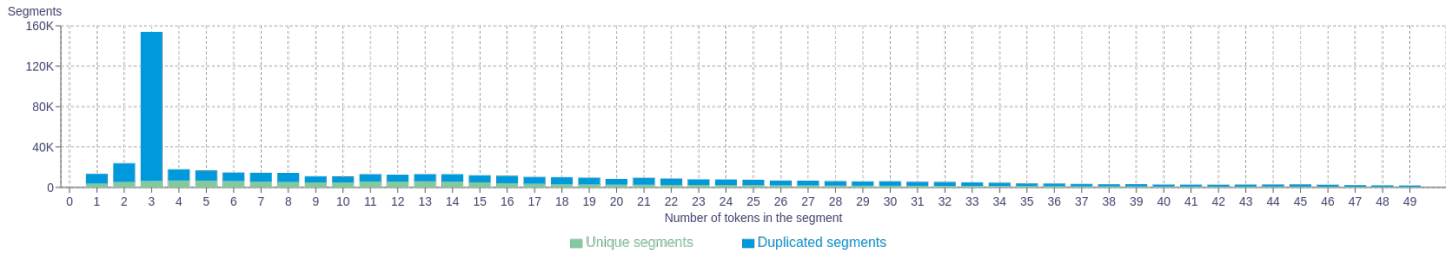
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (11K documents)

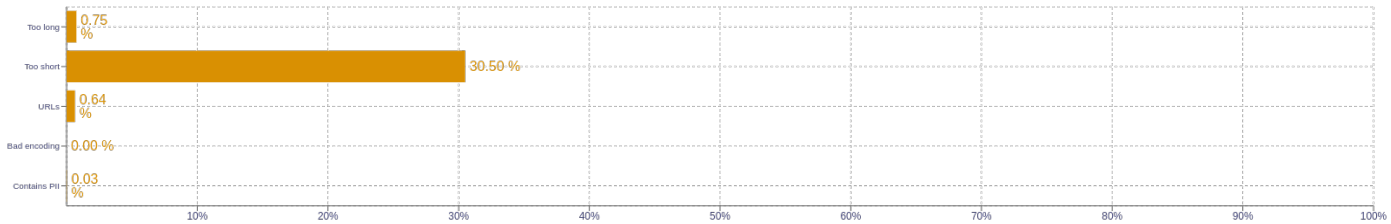


Segment length distribution by token

<= 49 tokens = **141K** segments | **402K** duplicates
> 50 tokens = **59K** segments | **38K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	in | 303017 · balinese | 99435 · indonesian | 97372 · english | 97267 · bali | 80618
2	in balinese | 97070 · in english | 96934 · in indonesian | 96516 · hyang widi | 12563 · widi wasa | 12169
3	hyang widi wasa | 12122 · usage examples pulled | 1971 · pulled from the | 1971 · examples pulled from | 1971 · duwur nyane susunin | 1730
4	usage examples pulled from | 1971 · examples pulled from the | 1971 · pulled from the virtual | 1323 · from the virtual library | 1318 · susunin antuk aled sesayut | 1244
5	usage examples pulled from the | 1971 · examples pulled from the virtual | 1323 · pulled from the virtual library | 1318 · nyane susunin antuk aled sesayut | 1237 · duwur nyane susunin antuk aled | 1237

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