HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
bjn_Latn.jsonl.tsv	11/27/2024	Banjar (bjn)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
18,764	366,336	154,702 (42.23 %)	10M	53.33 MB	55,627,091

Top 10 domains

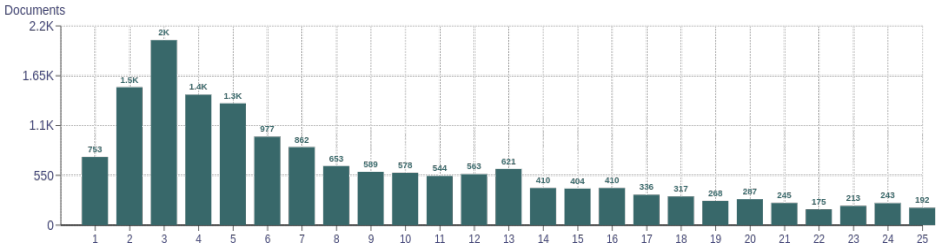
Domain	Docs	% of total
wikipedia.org	7.1K	37.59
blogspot.com	1.4K	7.62
wordpress.com	988	5.27
banjarmasinbungas.com	637	3.39
bible.is	535	2.85
sciencegraph.net	397	2.12
petalokasi.org	244	1.30
tribunnews.com	225	1.20
forvo.com	180	0.96
blogspot.co.id	167	0.89

Top 10 TLDs

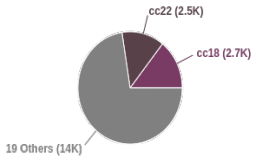
Domain	Docs	% of total
org	7.6K	40.67
com	7.4K	39.68
net	721	3.84
is	535	2.85
co.id	462	2.46
ac.id	292	1.56
go.id	181	0.96
id	177	0.94
asia	164	0.87
info	139	0.74

Documents size (in segments)

<= 25 segments **85.23%** (16K documents)
> 25 segments **14.77%** (2.8K documents)

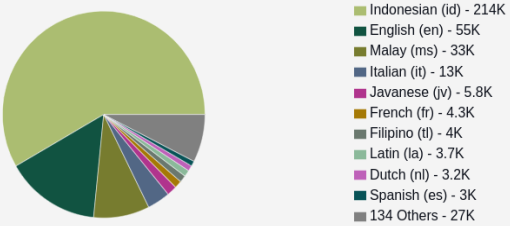


Documents by collection



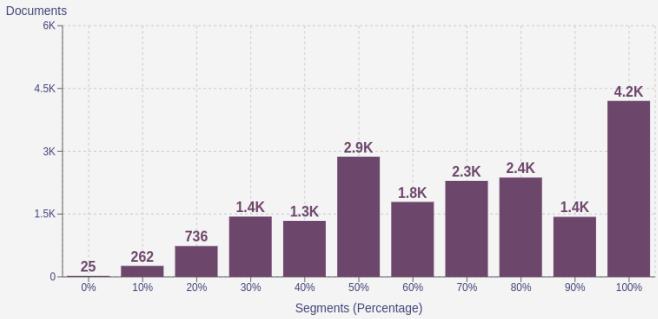
Language Distribution

Number of segments



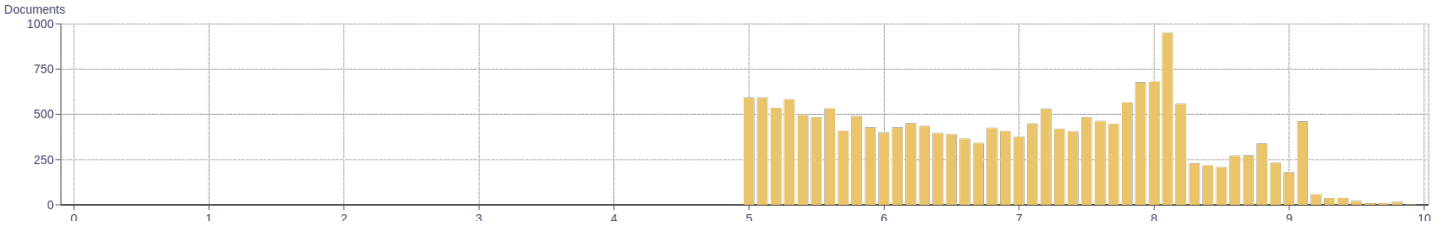
*Banjar (bjn) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Banjar (bjn) inside documents



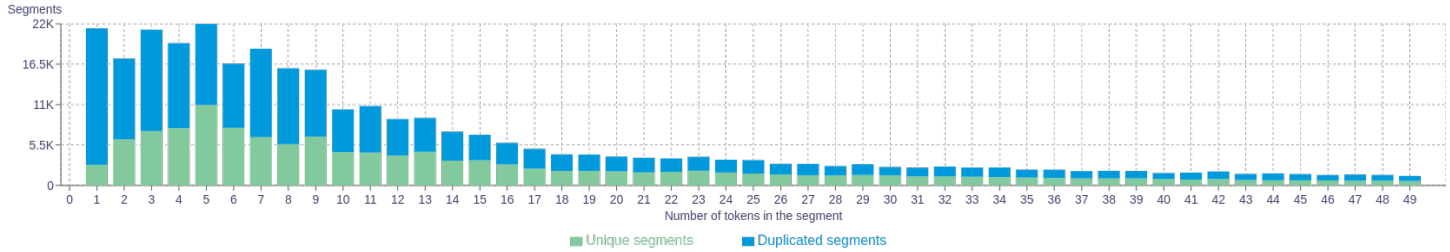
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (19K documents)

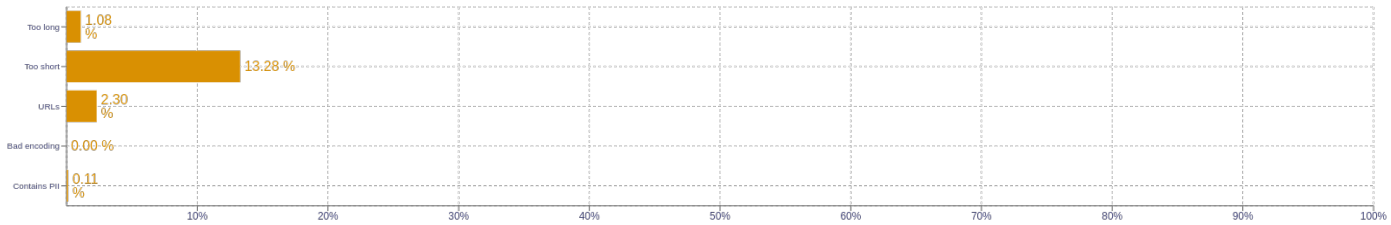


Segment length distribution by token

<= 49 tokens = **130K** segments | **182K** duplicates
> 50 tokens = **54K** segments | **29K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	arjuna | 28581 · banjarmasin | 28345 · hotel | 20187 · je | 19277 · caps | 18504
2	warfarin arjuna | 8435 · caps arjuna | 8399 · sunting sumber | 8337 · caps warfarin | 4921 · warfarin warfarin | 4891
3	caps warfarin arjuna | 2411 · warfarin warfarin arjuna | 2309 · jual rok tutu | 1714 · tutu di banjarmasin | 1712 · caps arjuna warfarin | 1677
4	rok tutu di banjarmasin | 1712 · grosir jual rok tutu | 1284 · tutu di banjarmasin murah | 1070 · caps warfarin arjuna warfarin | 846 · caps arjuna warfarin arjuna | 833
5	jual rok tutu di banjarmasin | 1712 · rok tutu di banjarmasin murah | 1070 · hotel hotel hotel hotel hotel | 810 · caps warfarin arjuna warfarin arjuna | 410 · warfarin warfarin arjuna warfarin arjuna | 400

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