HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-jav_Latn	9/17/2025	Javanese

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
239,461	6,079,314	4,059,403 (66.77 %)	165M	899,652,147	868.72 MB

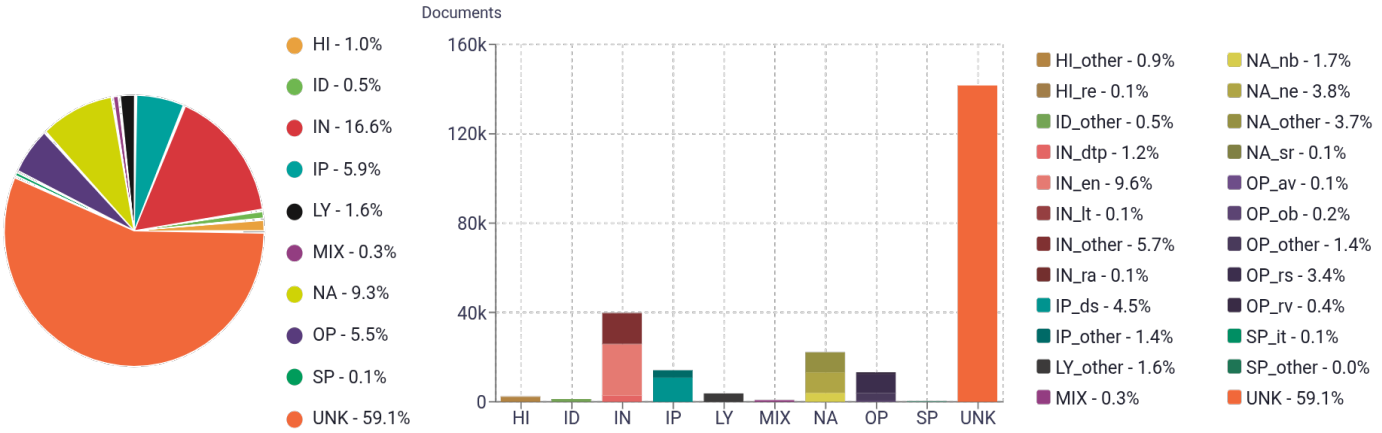
Top 10 domains

Domain	Docs	% of total
wikipedia.org	23K	9.76%
topwar.ru	7.2K	3.02%
eturbonews.com	5.7K	2.36%
bisnislink.com	5.4K	2.27%
blogspot.com	5.3K	2.23%
busanaarafah.com	4.4K	1.83%
wordpress.com	4.3K	1.79%
wondershare.com	4.1K	1.71%
martech.zone	3.9K	1.62%
wikisource.org	2.2K	0.93%

Top 10 TLDs

Domain	Docs	% of total
com	143K	59.61%
org	34K	14.29%
ru	8.4K	3.50%
net	7.3K	3.06%
icu	5.8K	2.41%
zone	3.9K	1.62%
co.id	3.8K	1.57%
xyz	2.9K	1.20%
top	1.8K	0.76%
news	1.8K	0.74%

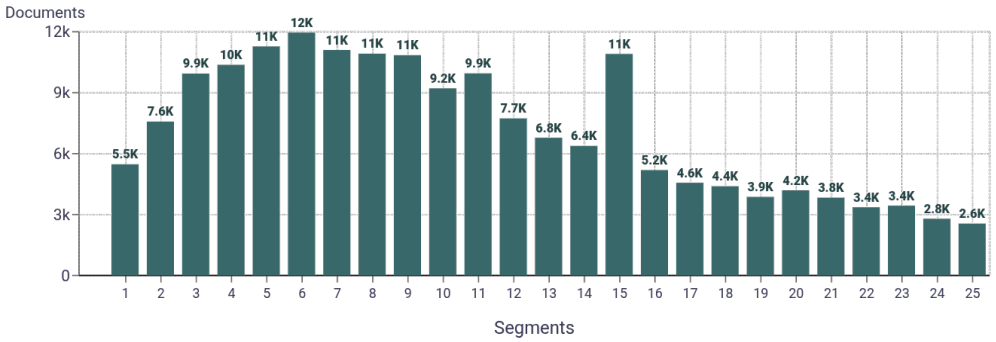
Register labels



MT:55.7% | 133K Documents

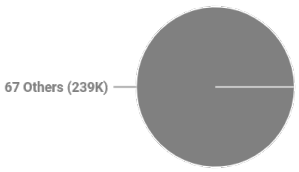
Documents size (in segments)

≤ 25 segments **74.55%** (179K documents)
> 25 segments **25.45%** (61K documents)



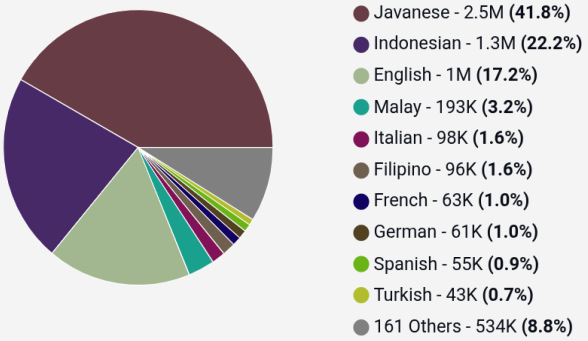
Document collections

CC = 90.04%
IA = 9.96%

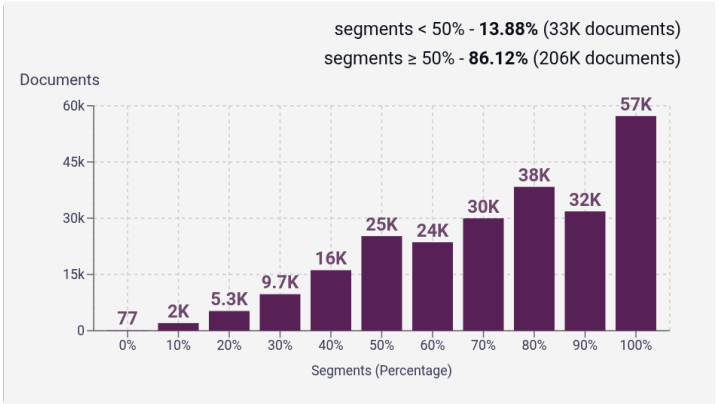


Language Distribution

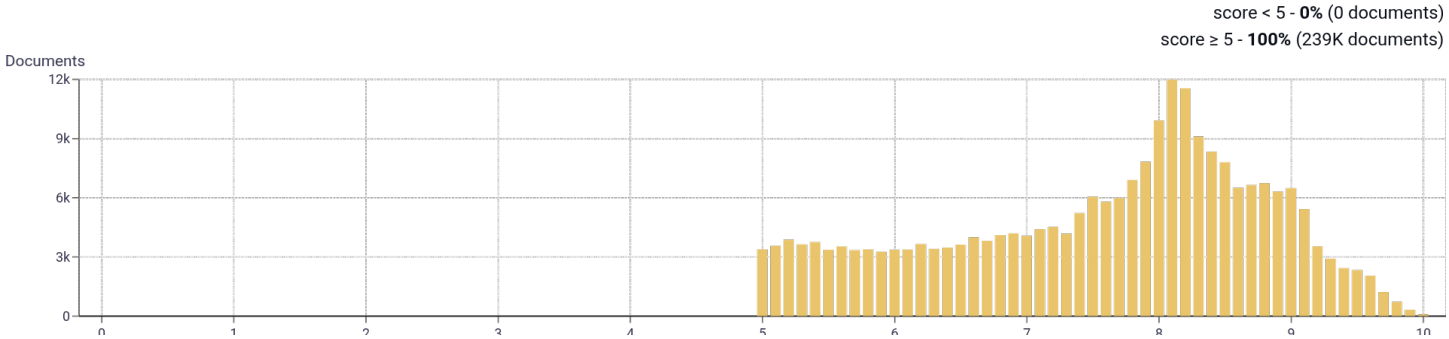
Number of segments in the Javanese corpus



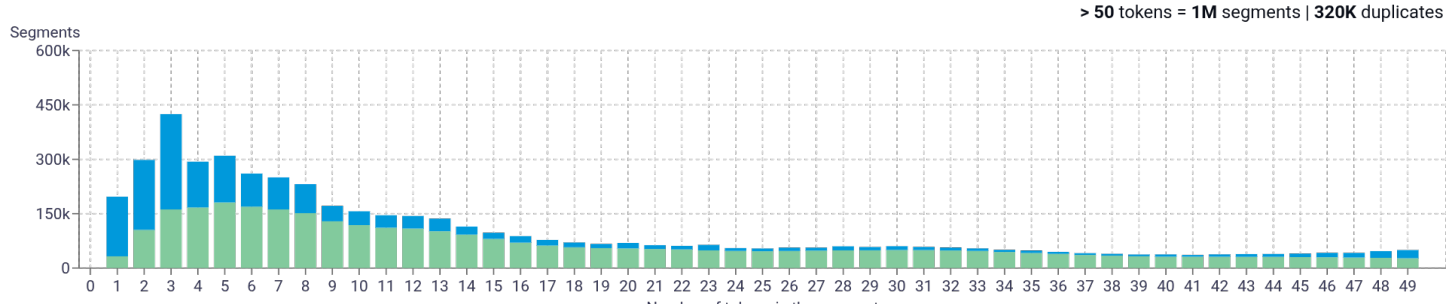
Percentage of segments in Javanese inside documents



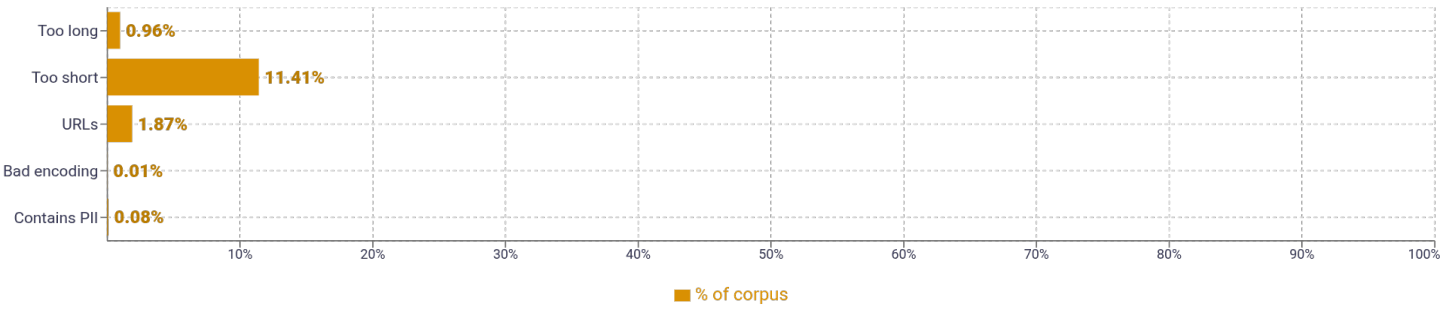
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	sing | 2,253,469 · ingkang | 517,364 · basa | 438,169 · jawa | 427,963 · piala | 323,356
2	piala donya | 203,709 · piala dunia | 94,236 · basa jawa | 57,833 · gusti allah | 44,484 · salah sawijining | 37,793
3	tohan maén bal | 34,743 · piala donya qatar | 30,281 · piala donya fifa | 17,300 · totoan piala donya | 16,563 · prediksi asil piala | 14,969
4	tohan maén bal online | 13,320 · situs tohan maén bal | 12,477 · sepak bola piala dunia | 9,120 · situs tohan piala dunya | 9,069 · prediksi asil piala dunia | 8,486
5	platform tohan maén bal bébas | 6,738 · situs tohan maén bal pangalusna | 6,368 · situs tohan maén bal légal | 6,109 · tim sepak bola piala dunia | 4,585 · naon situs tohan piala dunya | 3,867

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

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	**HI**	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	**Informational persuasion**	**IP**	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	**Opinion**	**OP**
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	**Informational description**	**IN**	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				