

General overview

Corpus	Date	Language
jav_Latn.jsonl.tsv	9/24/2024	Javanese (jv)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
195,966	6,430,750	2,773,932 (43.14 %)	170M	931,278,526	903.28 MB

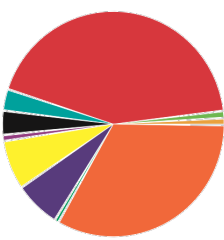
Top 10 domains

Domain	Docs	% of total
wikipedia.org	76K	38.70%
blogspot.com	11K	5.61%
wordpress.com	6.7K	3.41%
busanaarafah.com	3.1K	1.58%
sastra.org	2.5K	1.27%
bisnislink.com	2.1K	1.09%
blogspot.co.id	2K	1.01%
eturbonews.com	1.8K	0.93%
topwar.ru	1.6K	0.81%
expertpokupay.news	1.4K	0.69%

Top 10 TLDs

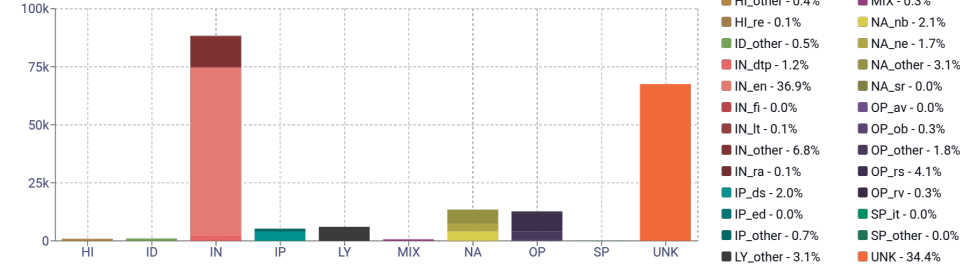
Domain	Docs	% of total
org	85K	43.48%
com	78K	39.94%
net	4.6K	2.34%
co.id	4.4K	2.23%
icu	3.5K	1.78%
ru	2K	1.03%
news	1.5K	0.76%
info	1.4K	0.73%
top	1.1K	0.55%
web.id	902	0.46%

Register labels

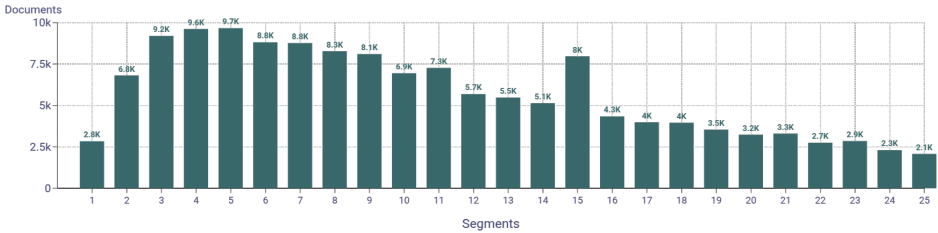


MT:30.1% | 59K Documents

Documents

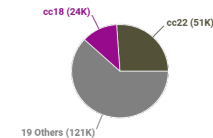


Documents size (in segments)



<= 25 segments 72.9% (143K documents)
> 25 segments 27.1% (53K documents)

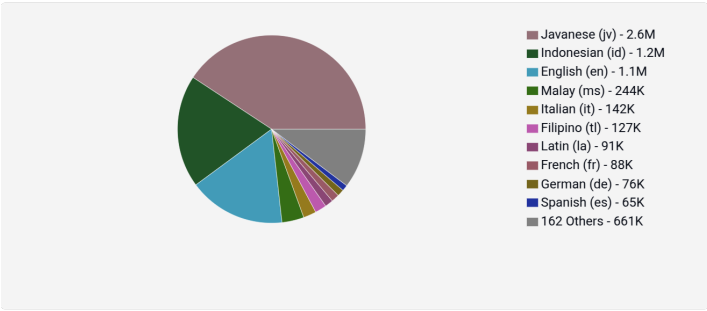
Documents by collection



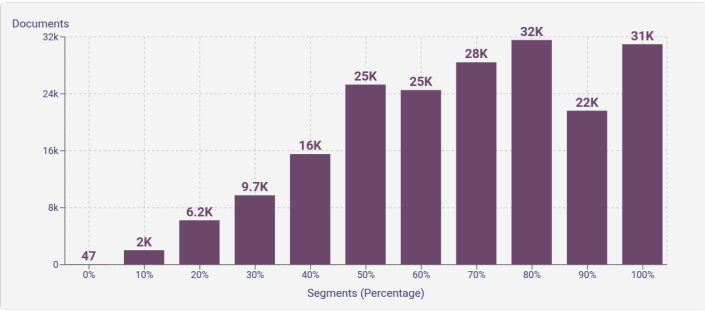
CC = 66.69%
IA = 33.31%

Language Distribution

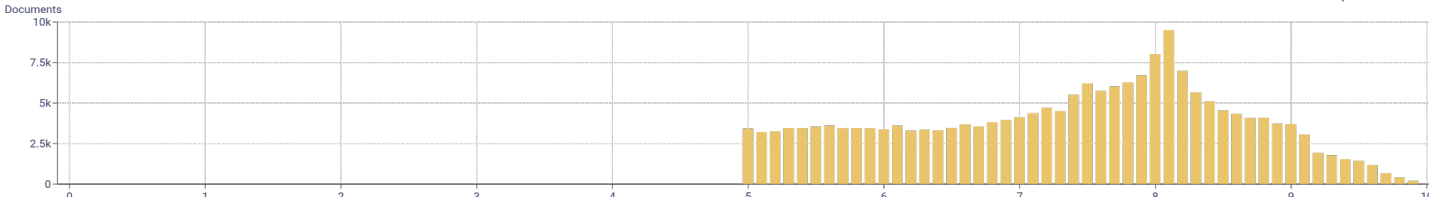
Number of segments in the Javanese (jv) corpus



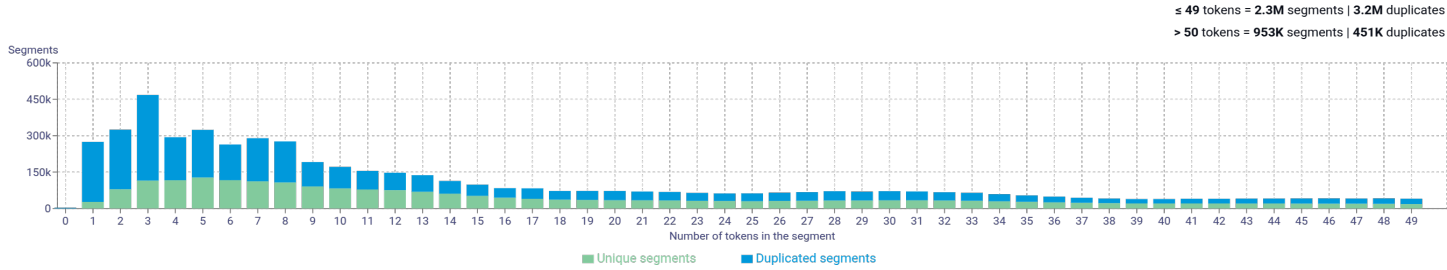
Percentage of segments in Javanese (jv) inside documents



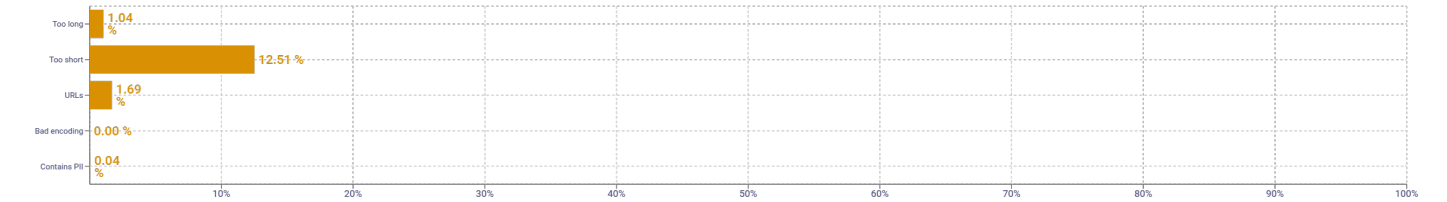
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	sing 1615589 ingkang 940452 punika 466482 kanthi 394011 saking 295739
2	piala donya 121060 besut sumber 87733 suntung sumber 82631 inggih punika 70405 piala dunia 52873
3	tohan maén bal 20531 piala donya qatar 17602 b c d 14312 c d e 10964 bab lan paragraf 10363
4	b c d e 10960 c d e f 8487 situs tohan maén bal 7598 tohan maén bal online 7393 d e f g 6735
5	b c d e f 8486 c d e f g 6735 d e f g h 5348 e f g h i 4401 platform tohan maén bal bébas 4195

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/pablopt16n/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				