HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
sun_Latn.jsonl.tsv	9/23/2024	Sundanese (su)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
114,755	3,237,757	1,564,280 (48.31 %)	86M	472,202,737	457.68 MB

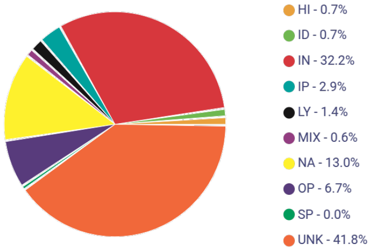
Top 10 domains

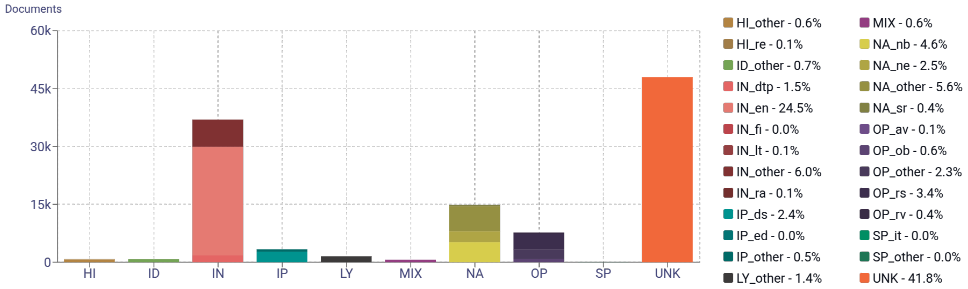
Domain	Docs	% of total
wikipedia.org	29K	25.24%
wordpress.com	7.9K	6.86%
blogspot.com	7.3K	6.32%
sundanet.com	2.8K	2.40%
mangle-online.com	1.4K	1.25%
sundanews.com	1.3K	1.13%
blogspot.co.id	1.3K	1.10%
fikminsunda.com	1.1K	0.94%
martech.zone	1.1K	0.93%
bejatikoran.com	757	0.66%

Top 10 TLDs

Domain	Docs	% of total
com	62K	54.02%
org	32K	28.03%
icu	2.8K	2.41%
net	2.6K	2.24%
co.id	1.6K	1.36%
zone	1.1K	0.94%
top	894	0.78%
info	862	0.75%
is	713	0.62%
web.id	517	0.45%

Register labels

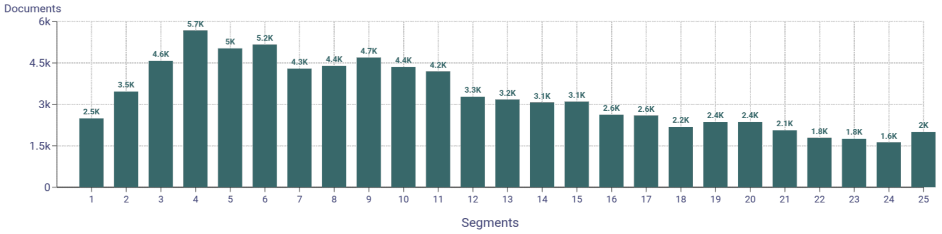




🤖 **MT**:36.6% | 42K Documents

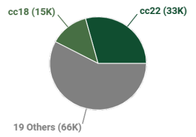
Documents size (in segments)

<= 25 segments **71.73%** (82K documents)
> 25 segments **28.27%** (32K documents)



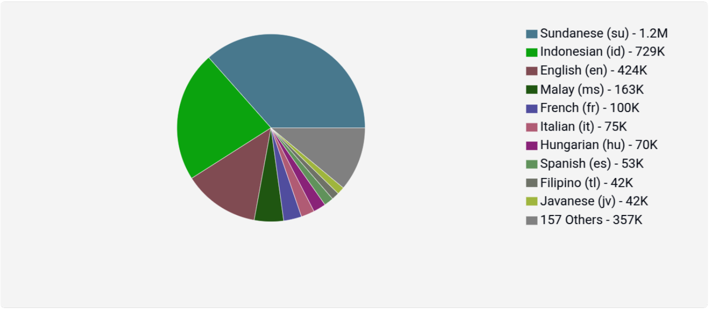
Documents by collection

CC = 68.67%
IA = 31.33%

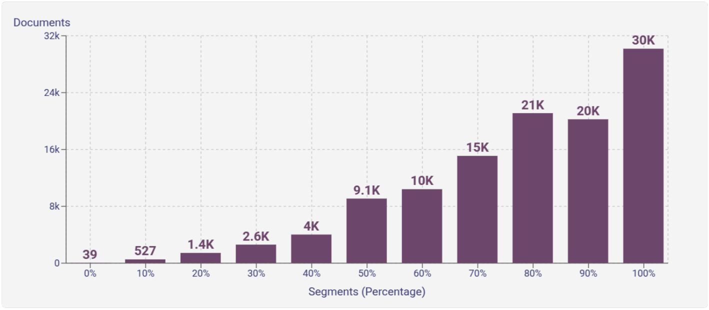


Language Distribution

Number of segments in the Sundanese (su) corpus

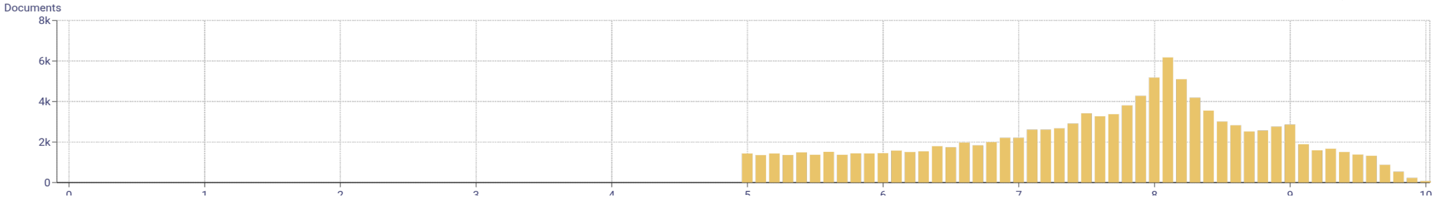


Percentage of segments in Sundanese (su) inside documents



Distribution of documents by document score

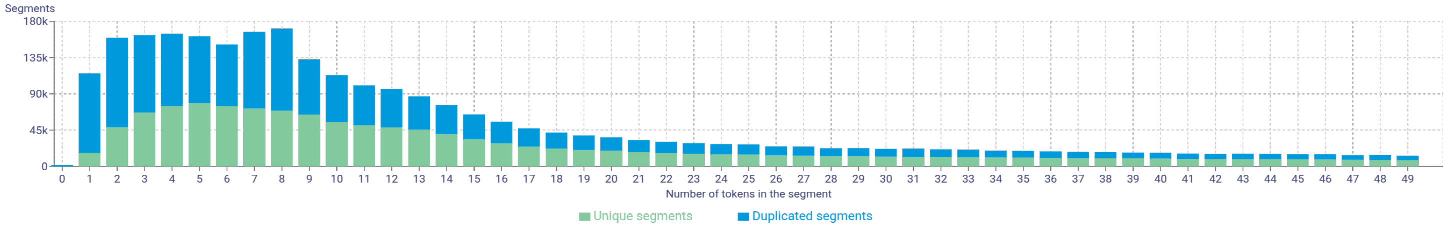
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (115K documents)



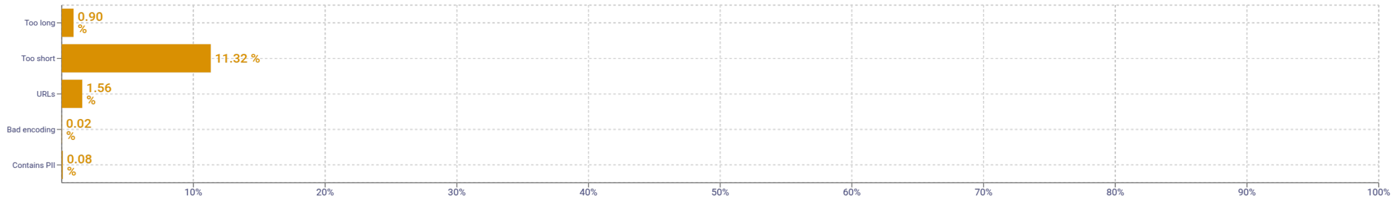
Segment length distribution by token

≤ 49 tokens = **1.3M** segments | **1.5M** duplicates

> 50 tokens = **505K** segments | **223K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	nu | 925069 ka | 683056 sareng | 472304 teu | 422505 ti | 389533
2	piala dunya | 93086 nepi ka | 71312 édit sumber | 47284 piala dunia | 43158 maén bal | 42615
3	tohan maén bal | 8926 skor piala dunya | 8131 piala dunya qatar | 7916 tohan piala dunya | 6691 maén bal online | 5932
4	usaha my healthy yoghurt | 4390 mitra usaha my healthy | 4390 sepak bola piala dunya | 3791 b c d e | 3630 skor piala dunya internasional | 3628
5	mitra usaha my healthy yoghurt | 4390 harga sepak bola piala dunya | 3336 b c d e f | 2881 hasil pencarian untuk kata kunci | 2560 bet dina maén bal online | 2313

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av