HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
ind_Latn.jsonl.tsv	7/26/2025	Indonesian (id)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
98,141,797	2,387,131,737	1,067,290,415 (44.71 %)	64B	381,930,808,361	357.31 GB

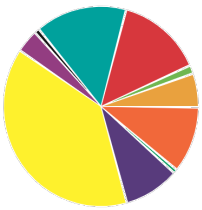
Top 10 domains

Domain	Docs	% of total
blogspot.com	8M	8.17%
wordpress.com	3.7M	3.77%
tribunnews.com	1.6M	1.68%
blogspot.co.id	1.3M	1.31%
kompas.com	783K	0.80%
blogspot.sg	470K	0.48%
us.com	406K	0.41%
detik.com	404K	0.41%
wikipedia.org	359K	0.37%
okezone.com	332K	0.34%

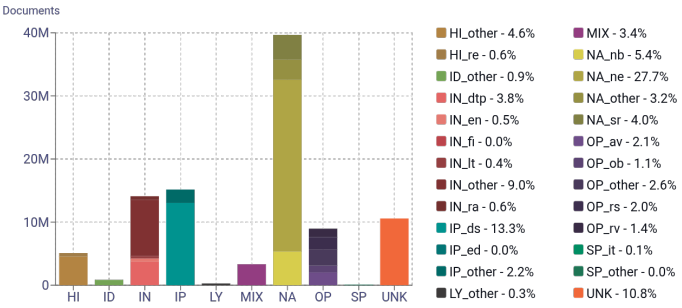
Top 10 TLDs

Domain	Docs	% of total
com	66M	66.77%
co.id	6M	6.07%
net	4M	4.09%
id	3.9M	4.02%
org	3.2M	3.24%
info	1.8M	1.85%
co	1.7M	1.72%
ac.id	1.6M	1.68%
go.id	1.5M	1.51%
web.id	1.1M	1.07%

Register labels



HI - 5.2%
ID - 0.9%
IN - 14.4%
IP - 15.4%
LY - 0.3%
MIX - 3.4%
NA - 40.4%
OP - 9.1%
SP - 0.1%
UNK - 10.8%

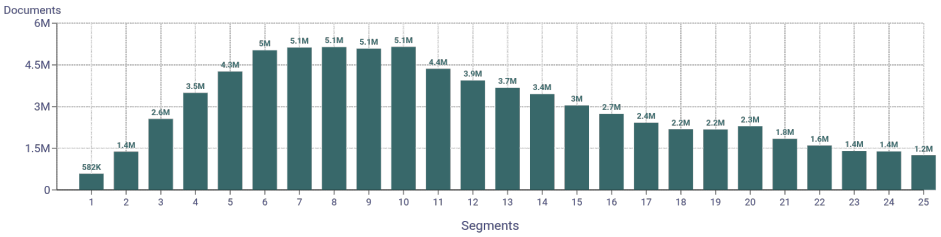


HI_other - 4.6%
HI_re - 0.6%
ID_other - 0.9%
IN_dtp - 3.8%
IN_en - 0.5%
IN_fi - 0.0%
IN_lt - 0.4%
IN_other - 9.0%
IN_ra - 0.6%
IP_ds - 13.3%
IP_ed - 0.0%
IP_other - 2.2%
LY_other - 0.3%
MIX - 3.4%
NA_nb - 5.4%
NA_ne - 27.7%
NA_other - 3.2%
NA_sr - 4.0%
OP_av - 2.1%
OP_ob - 1.1%
OP_other - 2.6%
OP_rs - 2.0%
OP_rv - 1.4%
SP_it - 0.1%
SP_other - 0.0%
UNK - 10.8%

MT:4.4% | 4.4M Documents

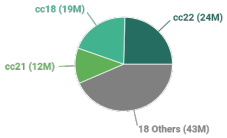
Documents size (in segments)

<= 25 segments **76.95%** (76M documents)
> 25 segments **23.05%** (23M documents)



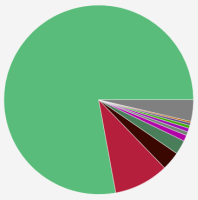
Documents by collection

CC = 71.47%
IA = 28.53%



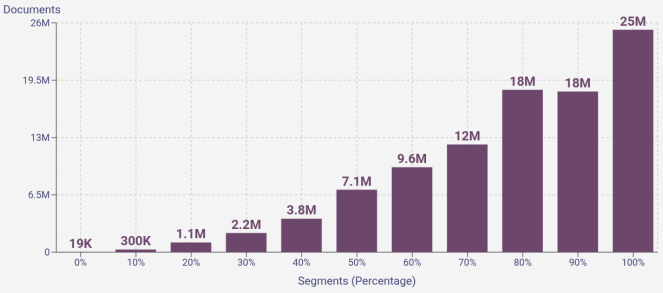
Language Distribution

Number of segments in the Indonesian (id) corpus



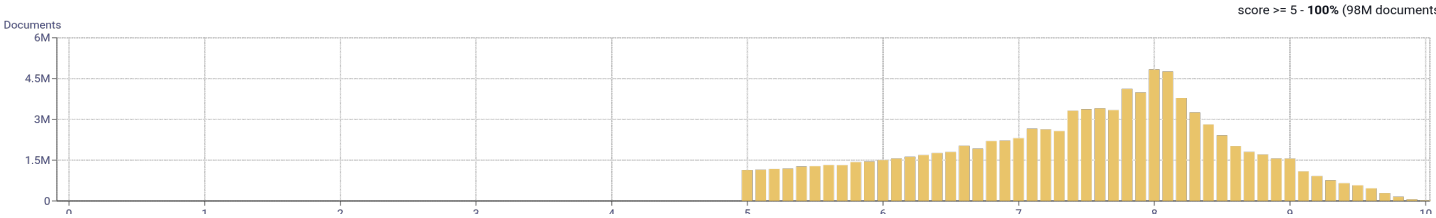
Indonesian (id) - 1.9B
English (en) - 223M
Malay (ms) - 76M
Italian (it) - 58M
French (fr) - 24M
Spanish (es) - 15M
German (de) - 15M
Dutch (nl) - 12M
Swedish (sv) - 9.8M
Esperanto (eo) - 6.8M
165 Others - 88M

Percentage of segments in Indonesian (id) inside documents



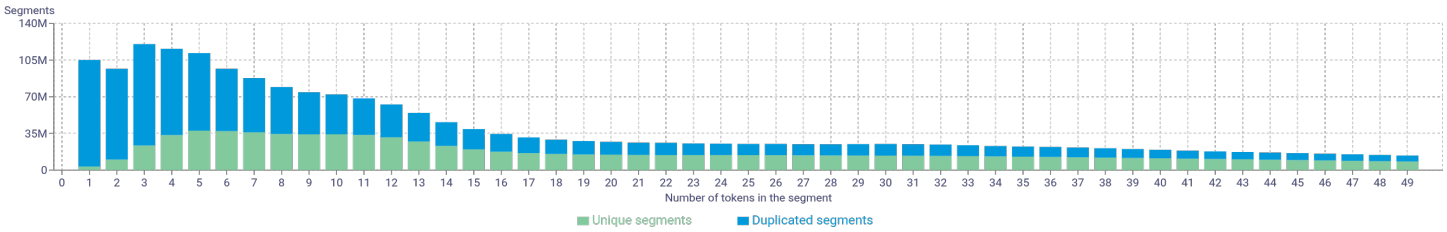
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (98M documents)

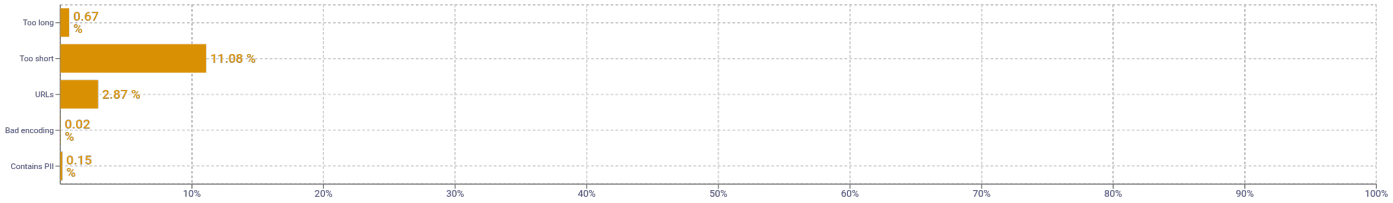


Segment length distribution by token

≤ 49 tokens = **2B** segments | **1.2B** duplicates
> 50 tokens = **374M** segments | **172M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	nya | 919650811 · orang | 134380517 · indonesia | 117256773 · memiliki | 100406959 · salah | 78546807
2	arti nya | 11130962 · orang tua | 10517856 · slot online | 8471962 · nama nya | 8036915 · judi online | 7947441
3	salah satu nya | 7814827 · bab i pendahuluan | 5658857 · judi slot online | 2631678 · tuhan yang maha | 2386592 · bolak balik sifon | 2335581
4	jilbab bolak balik sifon | 2335581 · tuhan yang maha esa | 2122539 · rahmat tuhan yang maha | 1689813 · bolak balik sifon polos | 1556883 · bab i pendahuluan a | 1484150
5	rahmat tuhan yang maha esa | 1636772 · jilbab bolak balik sifon polos | 1556883 · jilbab bolak balik pricilla warna | 811984 · jilbab bolak balik sifon motif | 778696 · elsa anna dan olaf ready | 731809

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