HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
zsm_Latn.jsonl.tsv	6/12/2025	Malay (zsm)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
18,416,407	579,115,786	186,999,394 (32.29 %)	14B	77,866,047,125	72.88 GB

Top 10 domains

Domain	Docs	% of total
blogspot.com	7.4M	40.40%
blogspot.my	1.1M	6.15%
wordpress.com	337K	1.83%
blogspot.sg	278K	1.51%
wikipedia.org	254K	1.38%
hotels.com	128K	0.69%
mstar.com.my	88K	0.48%
sinarharian.com.my	88K	0.48%
amanz.my	82K	0.44%
utusan.com.my	80K	0.44%

Top 10 TLDs

Domain	Docs	% of total
com	13M	69.96%
my	2M	10.68%
com.my	714K	3.88%
org	440K	2.39%
net	411K	2.23%
sg	303K	1.65%
gov.my	161K	0.87%
info	159K	0.86%
edu.my	86K	0.46%
com.au	52K	0.28%

Register labels

- HI - 3.8%
- ID - 1.3%
- IN - 9.2%
- IP - 8.6%
- LY - 0.1%
- MIX - 3.8%
- NA - 50.1%
- OP - 14.6%
- SP - 0.1%
- UNK - 8.3%

🤖 **MT**:4.1% | 755K Documents

- HI_other - 1.6%
- HI_re - 2.2%
- ID_other - 1.3%
- IN_dtp - 2.2%
- IN_en - 1.5%
- IN_fi - 0.0%
- IN_lt - 0.2%
- IN_other - 5.2%
- IN_ra - 0.1%
- IP_ds - 7.2%
- IP_ed - 0.0%
- IP_other - 1.4%
- LY_other - 0.1%
- MIX - 3.8%
- NA_nb - 21.8%
- NA_ne - 22.5%
- NA_other - 4.3%
- NA_sr - 1.4%
- OP_av - 0.7%
- OP_ob - 7.1%
- OP_other - 4.2%
- OP_rs - 2.0%
- OP_rv - 0.7%
- SP_it - 0.1%
- SP_other - 0.1%
- UNK - 8.3%
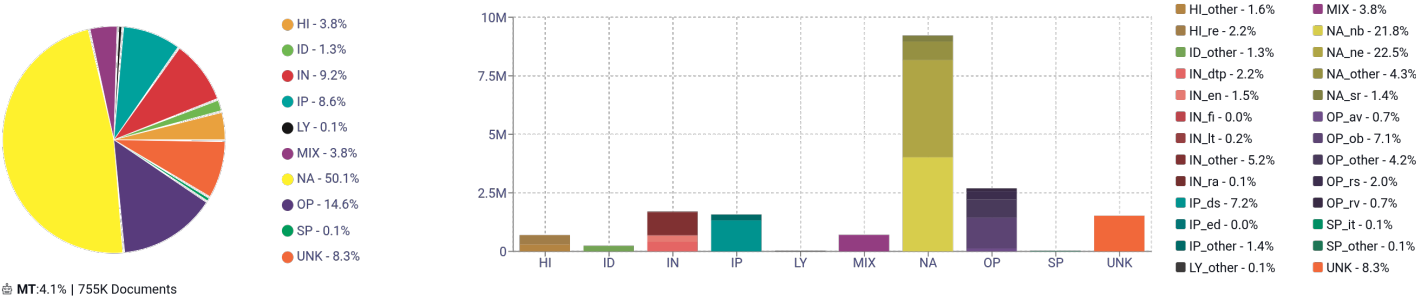
Documents size (in segments)
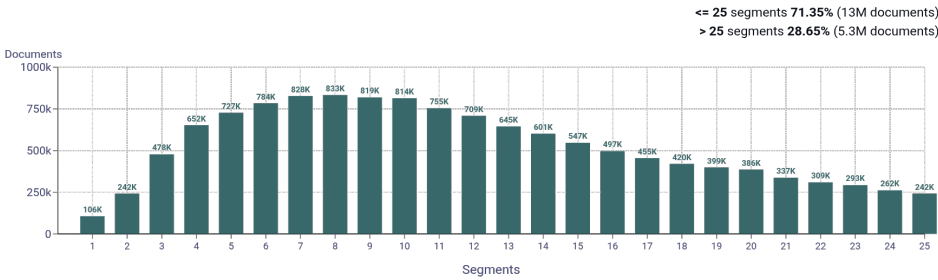
<= **25** segments **71.35%** (13M documents)
> **25** segments **28.65%** (5.3M documents)

Documents by collection

CC = 56.92%
IA = 43.08%

- cc22 (2.5M)
- wide15 (2.4M)
- cc18 (3.7M)
- 18 Others (9.8M)
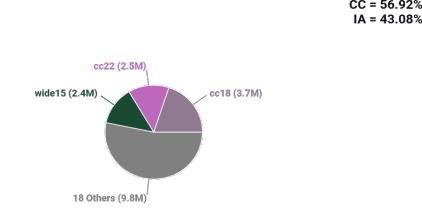
Language Distribution

Number of segments in the Malay (zsm) corpus

- Malay (zsm) - 361M
- English (en) - 83M
- Indonesian (id) - 60M
- Italian (it) - 19M
- Malay (ms) - 10M
- French (fr) - 6.4M
- Spanish (es) - 3.7M
- German (de) - 3.3M
- Dutch (nl) - 2.7M
- Turkish (tr) - 1.7M
- 166 Others - 27M
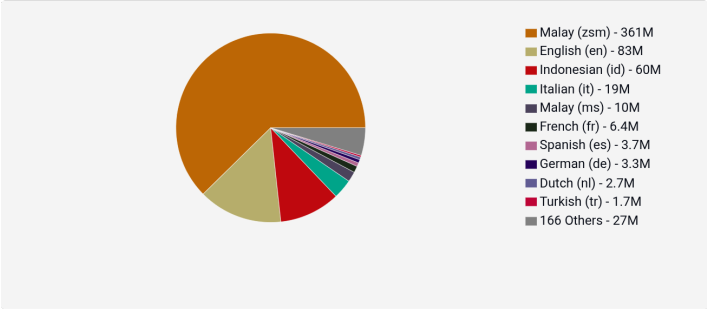
### Percentage of segments in Malay (zsm) inside documents
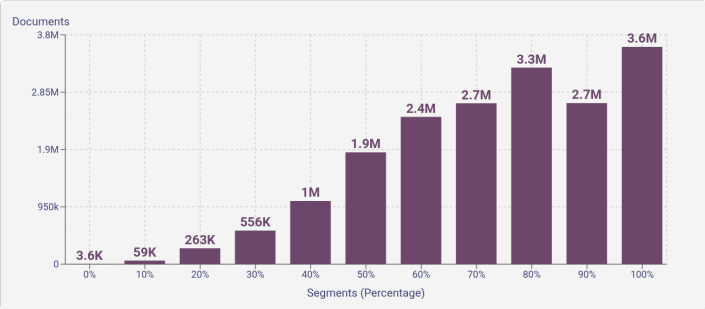
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (18M documents)

Segment length distribution by token

≤ 49 tokens = **159M** segments | **348M** duplicates
> 50 tokens = **72M** segments | **44M** duplicates

Segments

■ Unique segments ■ Duplicated segments

Number of tokens in the segment

Segment noise distribution

Category	Value
Too long	0.43 %
Too short	12.10 %
URLs	4.27 %
Bad encoding	0.01 %
Contains PII	0.18 %
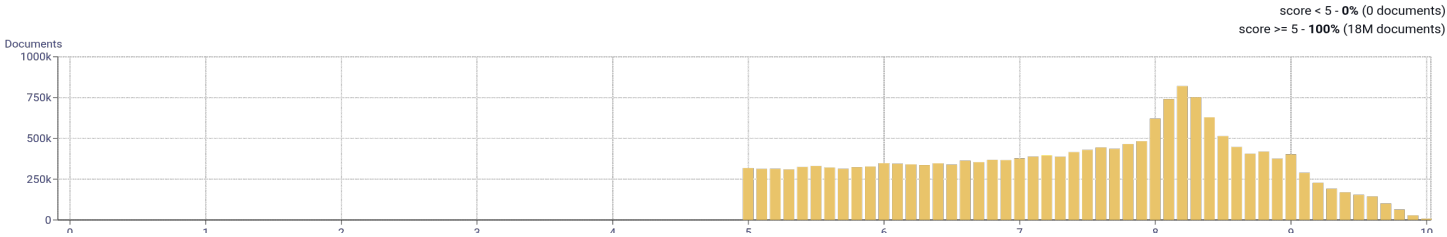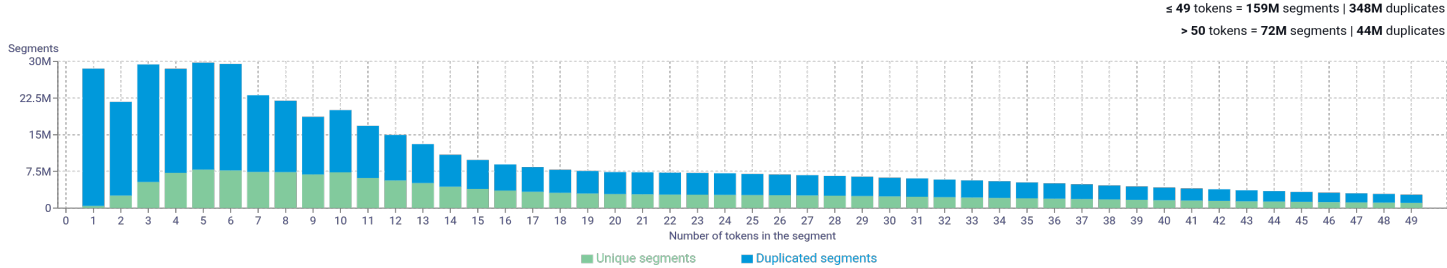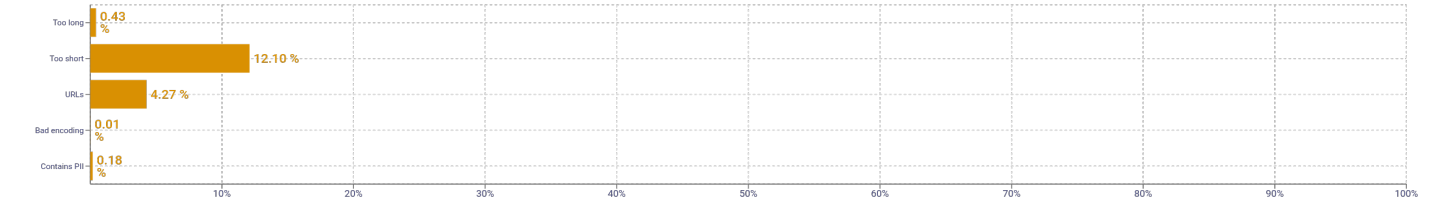
Frequent n-grams

Size	n-grams
1	aku | 50009310 tak | 45152232 nak | 39747769 dia | 38776599 ni | 37491506
2	kah kah | 3017683 terima kasih | 2874389 posted by | 2812286 aku tak | 1962356 allah swt | 1935757
3	kah kah kah | 1552130 to this post | 613259 links to this | 605921 sejam yang lalu | 578385 like or dislike | 390655
4	links to this post | 605812 comments links to this | 196416 nik abdul aziz nik | 171142 assalamualaikum dan salam sejahtera | 144987 kah kah kah kah | 136285
5	hotel ini dalam masa sejam | 571563 comments links to this post | 196414 nik abdul aziz nik mat | 169706 berkongsi permainan ini dengan kawan | 95959 jangan lupa untuk mengadar permainan | 94420

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				