HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
HPLT-v2-zsm_Latn.tsv	9/22/2024	Malay (zsm)

Volumes

Docs	Segments	Characters	Size
18,416,407	579,818,034	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

Pie chart:
- 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

Documents (bar chart by register):
- 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%
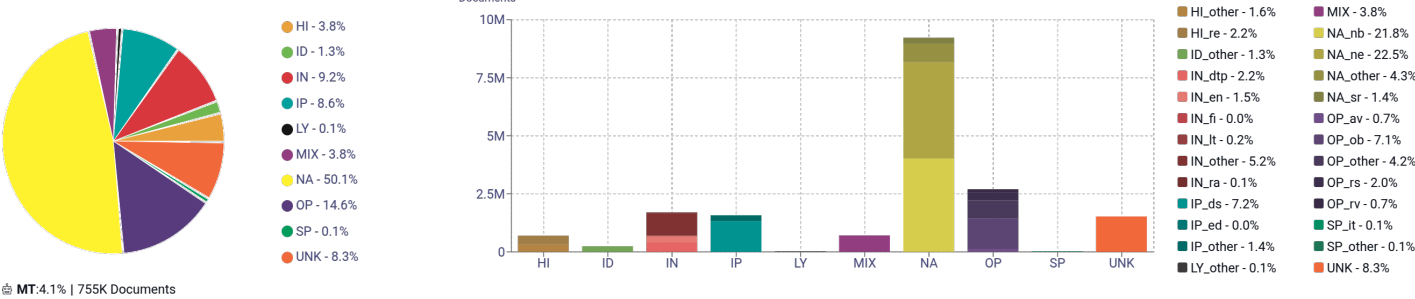
Documents size (in segments)

<= 25 segments **71.35%** (13M documents)
> 25 segments **28.65%** (5.3M documents)

Bar chart values (Segments: Documents):
1: 106K, 2: 242K, 3: 478K, 4: 652K, 5: 727K, 6: 784K, 7: 828K, 8: 833K, 9: 819K, 10: 814K, 11: 755K, 12: 709K, 13: 645K, 14: 601K, 15: 547K, 16: 497K, 17: 455K, 18: 420K, 19: 399K, 20: 386K, 21: 337K, 22: 309K, 23: 293K, 24: 262K, 25: 242K
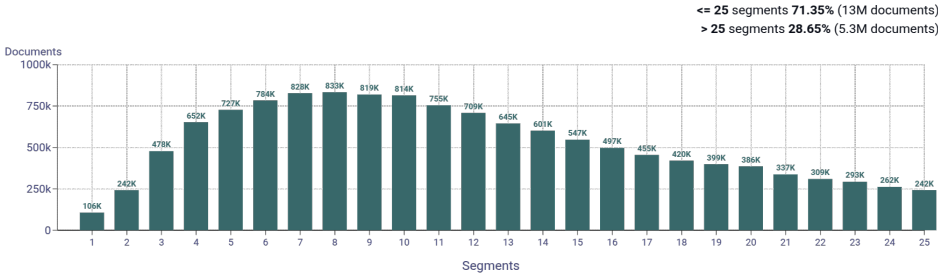
Documents by collection

CC = 56.92%
IA = 43.08%

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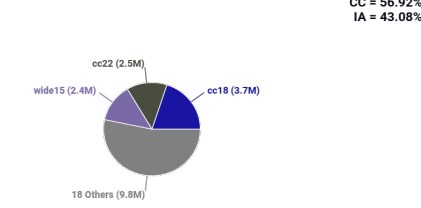
Language Distribution

Number of segments in the Malay (zsm) corpus

- Malay (zsm) - 361M
- English (en) - 84M
- 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
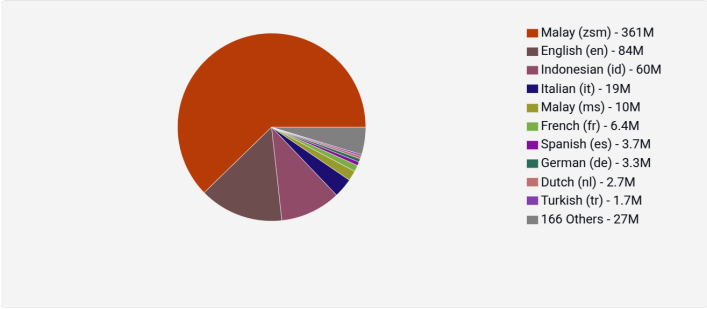
Percentage of segments in Malay (zsm) inside documents
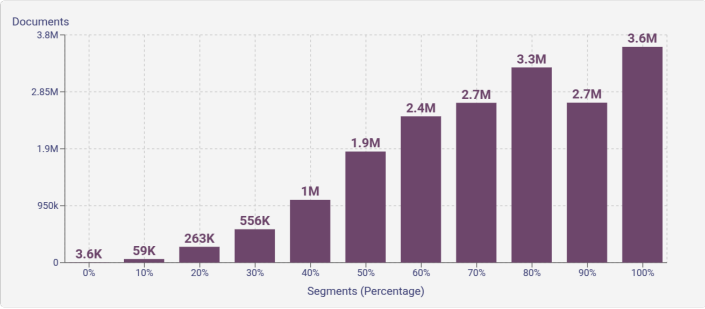
Documents (by Segments Percentage):
- 0%: 3.6K
- 10%: 59K
- 20%: 263K
- 30%: 556K
- 40%: 1M
- 50%: 1.9M
- 60%: 2.4M
- 70%: 2.7M
- 80%: 3.3M
- 90%: 2.7M
- 100%: 3.6M

Distribution of documents by document score

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

Segment noise distribution

- Too long: 0.43%
- Too short: 12.20%
- URLs: 4.27%
- Bad encoding: 0.01%
- Contains PII: 0.18%
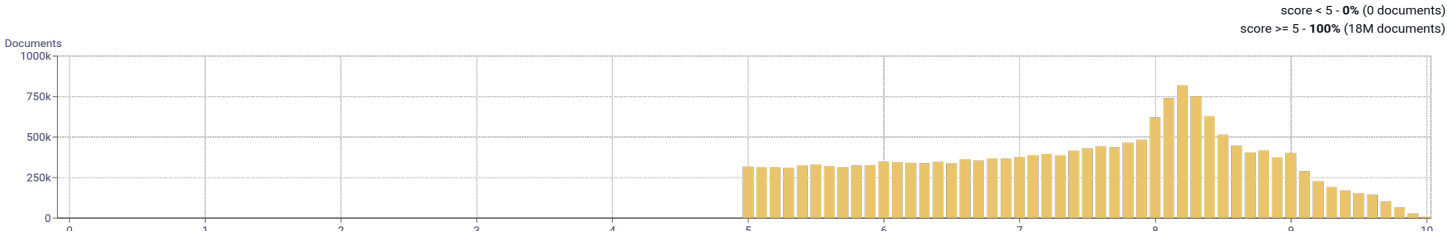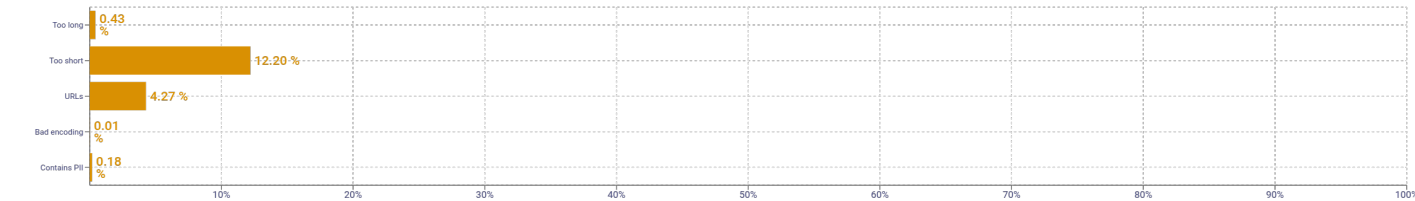
About HPLT Analytics

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