HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
tha_Thai.jsonl.tsv	6/14/2025	Thai (th)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
17,703,320	338,974,062	171,766,208 (50.67 %)	9.6B	59,654,618,398	152.11 GB

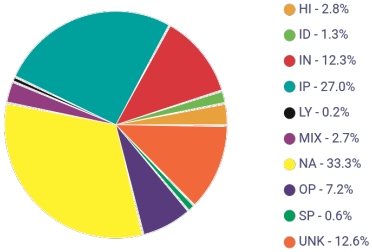
Top 10 domains

Domain	Docs	% of total
blogspot.com	330K	1.86%
sanook.com	255K	1.44%
wikipedia.org	229K	1.29%
mthai.com	221K	1.25%
thairath.co.th	194K	1.10%
tripadvisor.com	172K	0.97%
wordpress.com	150K	0.85%
plazathai.com	133K	0.75%
ryt9.com	129K	0.73%
newswit.com	119K	0.67%

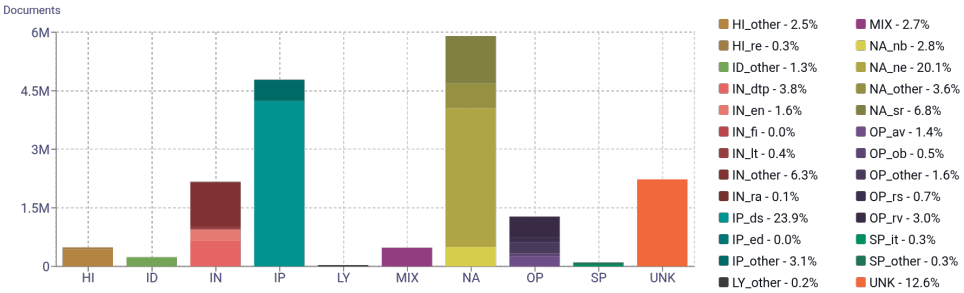
Top 10 TLDs

Domain	Docs	% of total
com	12M	69.22%
net	997K	5.63%
org	995K	5.62%
co.th	766K	4.33%
in.th	320K	1.81%
ac.th	220K	1.24%
co	217K	1.23%
go.th	198K	1.12%
or.th	191K	1.08%
tk	171K	0.97%

Register labels

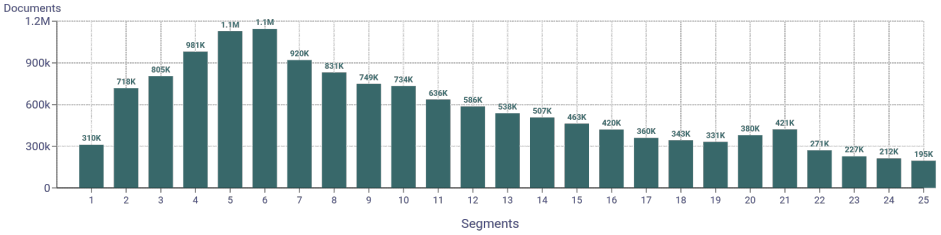


MT:6.9% | 1.2M Documents



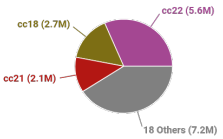
Documents size (in segments)

<= 25 segments **80.28%** (14M documents)
> 25 segments **19.72%** (3.5M documents)



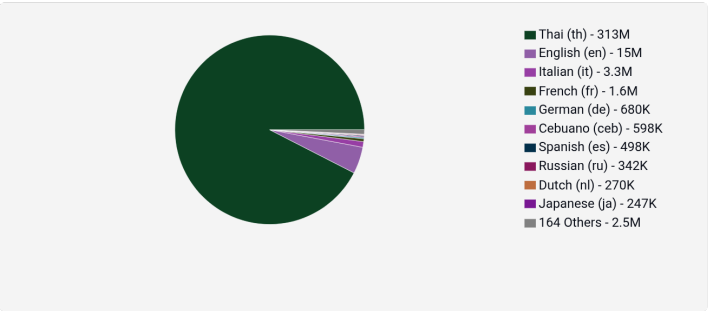
Documents by collection

CC = 69.10%
IA = 30.90%

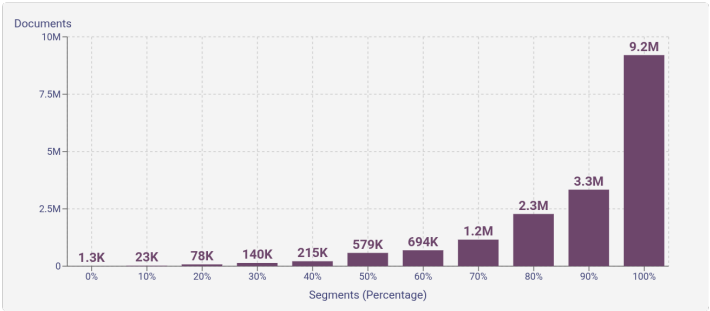


Language Distribution

Number of segments in the Thai (th) corpus

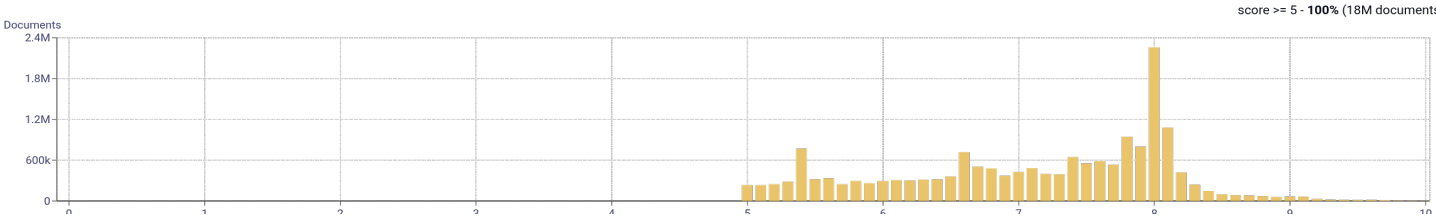


Percentage of segments in Thai (th) inside documents



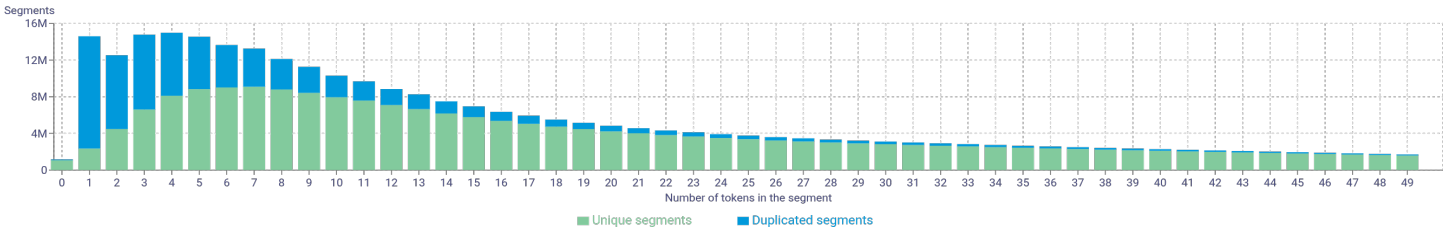
Distribution of documents by document score

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

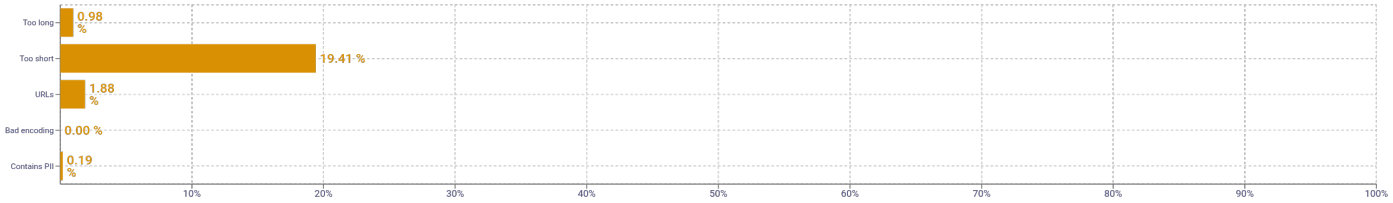


Segment length distribution by token

≤ 49 tokens = **206M** segments | **78M** duplicates
> 50 tokens = **55M** segments | **2.3M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	า | 42283047 · น | 34472796 · ก า | 33296705 · ๆ | 31824324 · ง | 31231038
2	ก า ร | 8478458 · ค ุ ณ | 8188478 · ค ว าม | 2815738 · ง า น | 2095526 · ม า | 2061429
3	o o o o | 360593 · ค า สิ โ น ออนไลน์ | 133824 · ค า ร์ า | 108670 · l l l l | 81011 · beauty clinic near me | 67765
4	o o o o o | 352884 · l l l l l | 79439 · บ า ค า ร์ | 77107 · ค า สิ โ น ออน ไลน์ | 51230 · ฝาก ถอน ไม่ มี ขั้น | 49204

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