HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-docslite.th.tsv	6/11/2024	Thai (th)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
8,192,709	605,713,959	165,778,486 (27.37 %)	11B	102.15 GB	

Top 10 domains

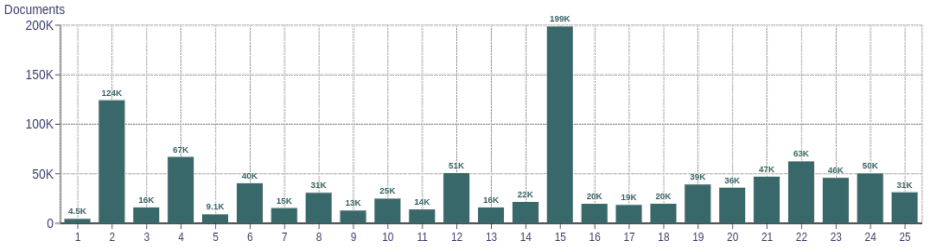
Domain	Docs	% of total
sinkardd.com	1.1M	12.82
thaimarketboard.com	778K	9.50
ddpromote.com	722K	8.81
thaibigplaza.com	676K	8.25
thai2plaza.com	387K	4.72
chanood.com	240K	2.93
diebuchsuche.com	124K	1.51
ju8.me	103K	1.26
ddhomeland.com	102K	1.25
blogspot.com	86K	1.05

Top 10 TLDs

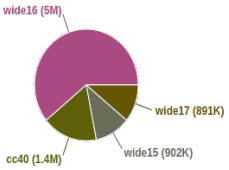
Domain	Docs	% of total
com	6.9M	83.87
net	274K	3.35
org	248K	3.03
me	115K	1.40
info	64K	0.79
in.th	55K	0.68
club	52K	0.63
co.th	50K	0.61
icu	38K	0.46
ac.th	34K	0.42

Documents size (in segments)

<= 25 segments **12.42%** (1M documents)
> 25 segments **87.58%** (7.2M documents)

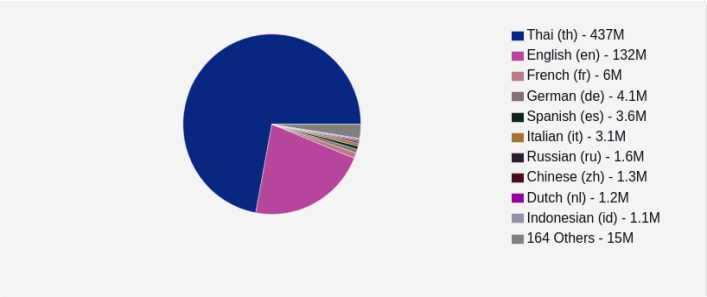


Documents by collection

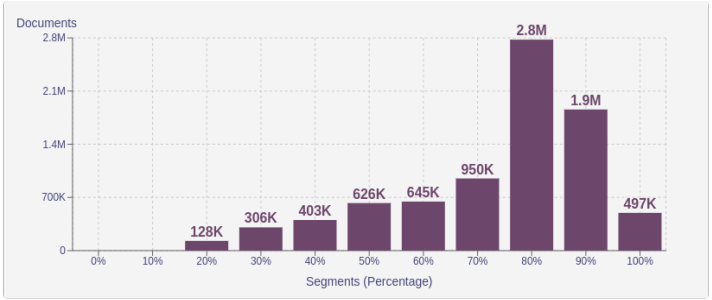


Language Distribution

Number of segments

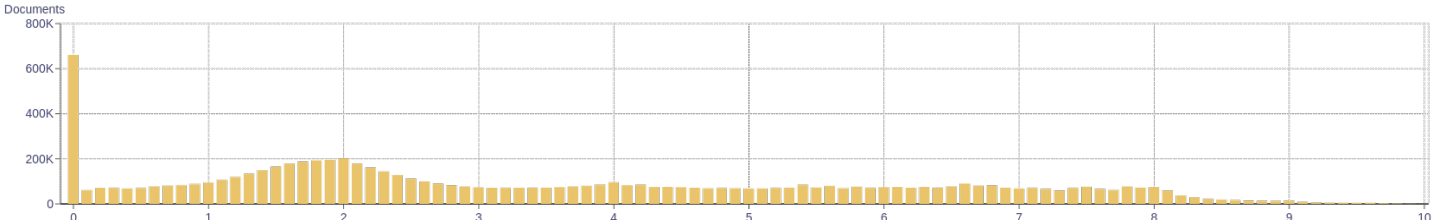


Percentage of segments in Thai (th) inside documents



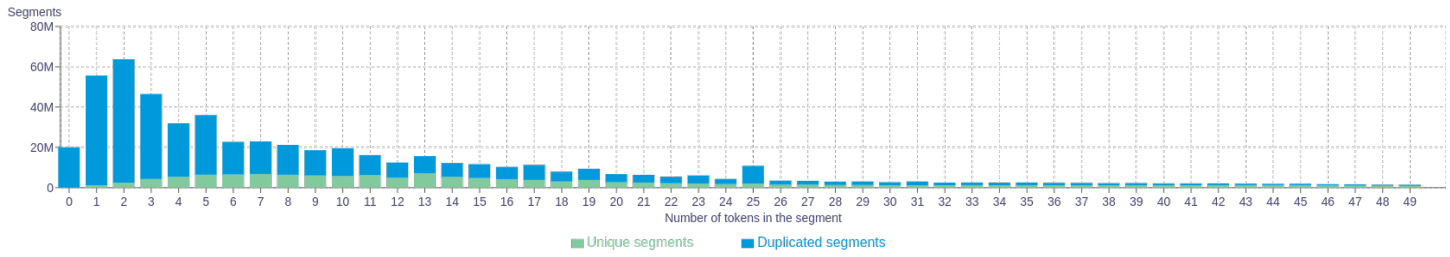
Distribution of documents by document score

score <= 5 - **68.88%** (5.6M documents)
score > 5 - **31.12%** (2.5M documents)



Segment length distribution by token

<= 49 tokens = **136M** segments | **426M** duplicates
> 50 tokens = **44M** segments | **14M** duplicates



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