HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
tam_Taml.jsonl.tsv	6/7/2025	Tamil (ta)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
6,105,834	168,490,322	66,803,246 (39.65 %)	3.6B	26,076,202,009	64.51 GB

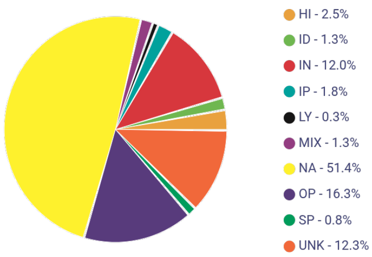
Top 10 domains

Domain	Docs	% of total
blogspot.com	522K	8.55%
blogspot.in	264K	4.32%
wikipedia.org	202K	3.30%
dinamalar.com	146K	2.39%
wordpress.com	105K	1.73%
vikatan.com	104K	1.70%
dinamani.com	93K	1.52%
blogspot.sg	86K	1.41%
dinakaran.com	84K	1.37%
blogspot.ch	83K	1.36%

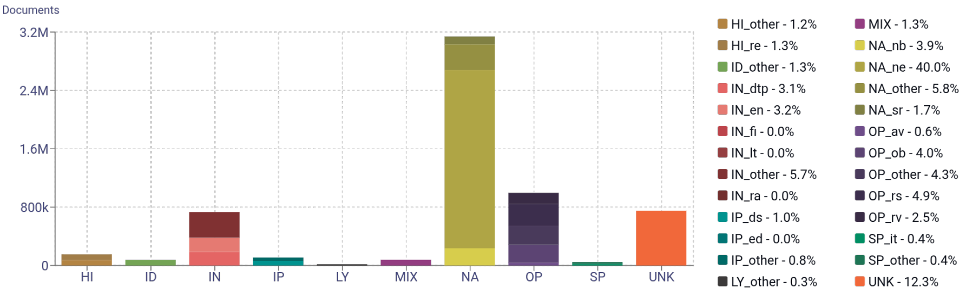
Top 10 TLDs

Domain	Docs	% of total
com	4.1M	66.66%
in	661K	10.82%
org	383K	6.27%
net	218K	3.57%
lk	119K	1.95%
sg	88K	1.44%
ch	86K	1.40%
info	45K	0.74%
ca	45K	0.74%
ae	39K	0.64%

Register labels

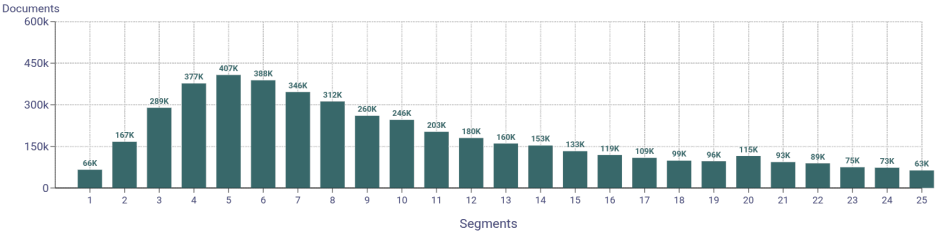


MT:2.4% | 144K Documents



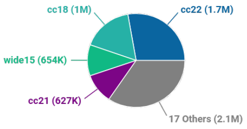
Documents size (in segments)

<= 25 segments **75.65%** (4.6M documents)
> 25 segments **24.35%** (1.5M documents)



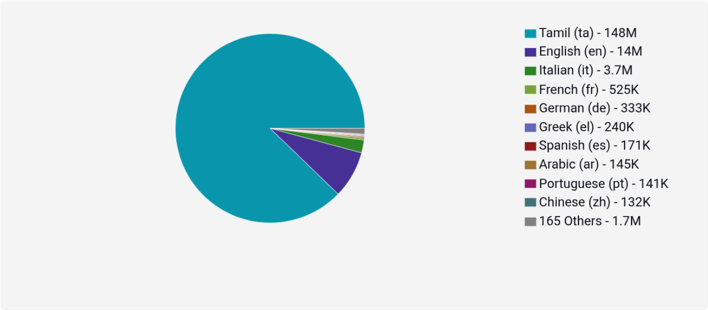
Documents by collection

CC = 63.99%
IA = 36.01%

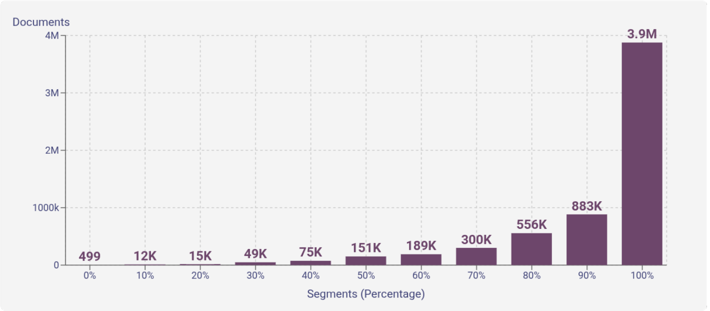


Language Distribution

Number of segments in the Tamil (ta) corpus

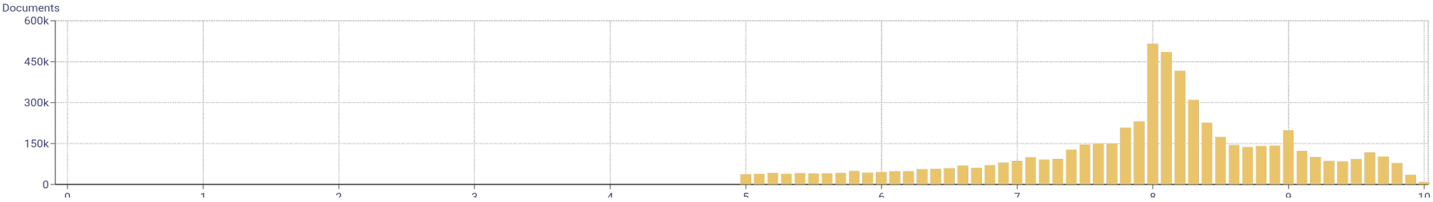


Percentage of segments in Tamil (ta) inside documents



Distribution of documents by document score

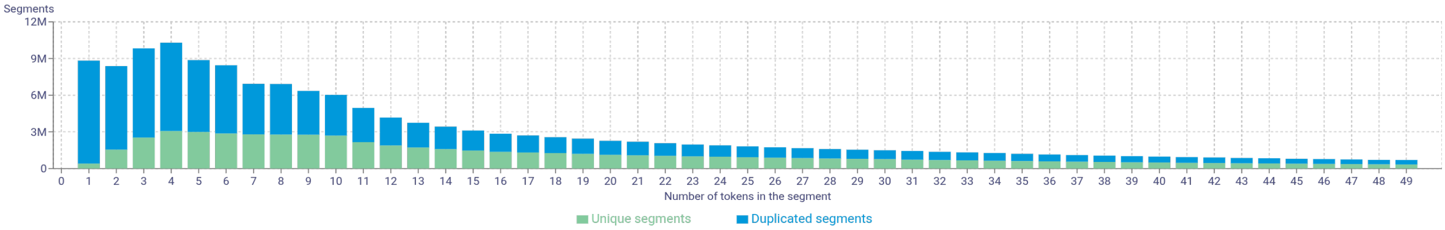
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (6.1M documents)



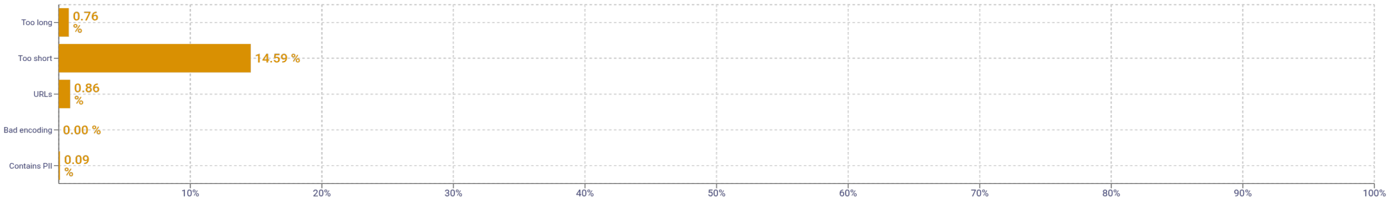
Segment length distribution by token

≤ 49 tokens = **58M** segments | **92M** duplicates

> 50 tokens = **18M** segments | **9.9M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ஆனால் | 6428658 இல்லை | 4712991 செய்து | 3973393 the | 3623648 தமிழ் | 3500763
2	read more | 474643 ஆம் ஆண்டு | 440567 posted by | 438100 of the | 422753 மத்திய அரசு | 354234
3	leave a comment | 126281 to this post | 98001 links to this | 97453 comments links to | 58395 தமிழ் ஓவியா said | 56859
4	links to this post | 97360 comments links to this | 58394 the rest of this | 39854 read the rest of | 39853 rest of this entry | 39844
5	-ஜட்ஜ் பலராமையா அவர்கள் கேள்வி பதில் | 23985 turned off anytime from browser | 20758 off anytime from browser settings | 20758 notifications can be turned off | 20758 can be turned off anytime | 20758

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