HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-mar_Deva	9/18/2025	Marathi (mr)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
6,460,836	81,800,481	61,631,308 (75.34 %)	2.8B	16,124,829,013	39.11 GB

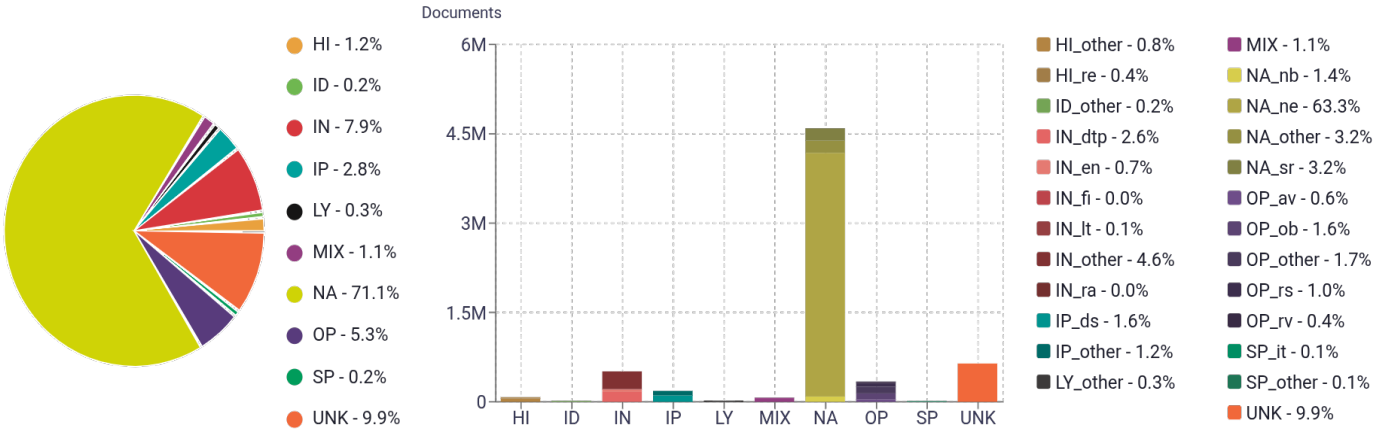
Top 10 domains

Domain	Docs	% of total
esakal.com	386K	5.97%
loksatta.com	301K	4.66%
bhaskar.com	265K	4.10%
news18.com	157K	2.43%
indiatimes.com	153K	2.37%
tv9marathi.com	117K	1.80%
lokmat.com	88K	1.36%
pudhari.news	85K	1.31%
maharashtratime...	79K	1.23%
dainikprabhat.com	78K	1.21%

Top 10 TLDs

Domain	Docs	% of total
com	4.9M	75.36%
in	881K	13.64%
org	161K	2.49%
news	144K	2.23%
co.in	84K	1.29%
net	54K	0.83%
es	37K	0.57%
tv	33K	0.51%
gov.in	31K	0.49%
live	25K	0.39%

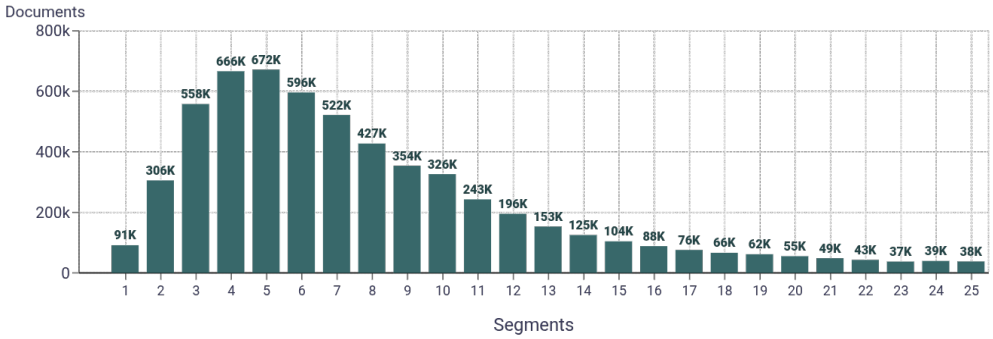
Register labels



MT:4.1% | 263K Documents

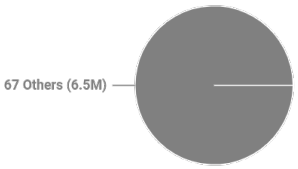
Documents size (in segments)

≤ 25 segments **91.24%** (5.9M documents)
> 25 segments **8.76%** (566K documents)



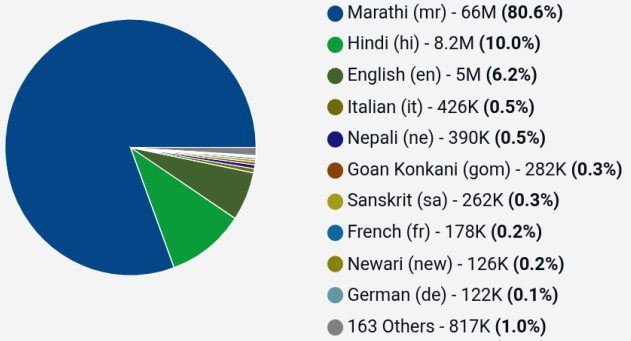
Document collections

CC = 96.22%
IA = 3.78%

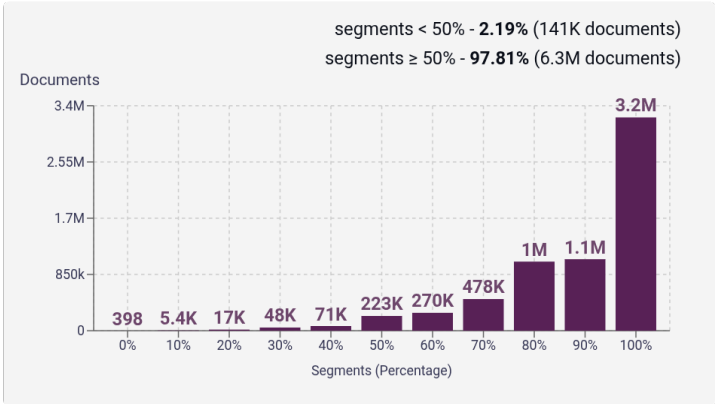


Language Distribution

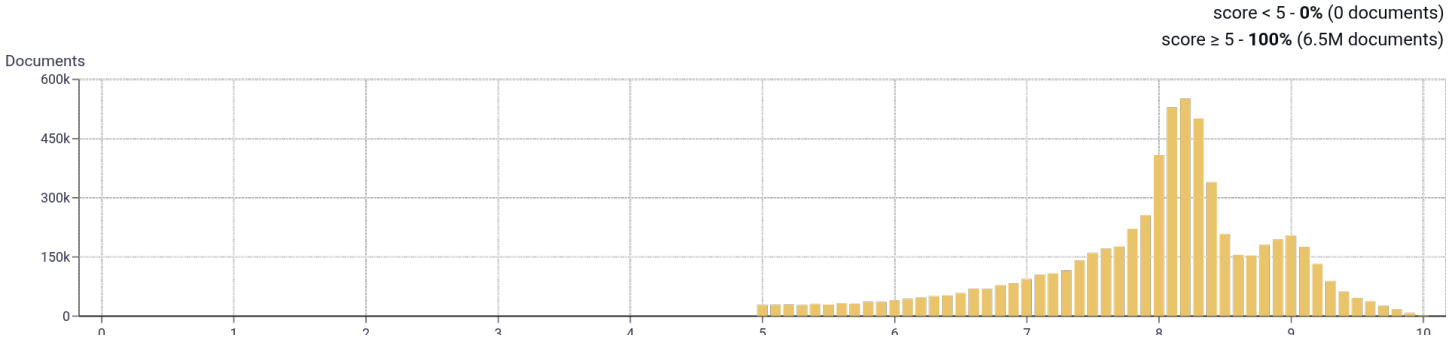
Number of segments in the Marathi (mr) corpus



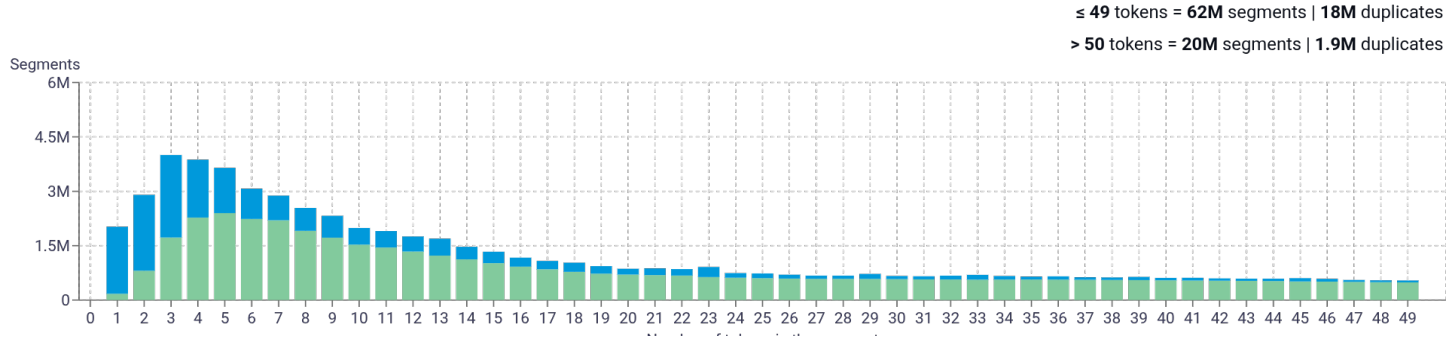
Percentage of segments in Marathi (mr) inside documents



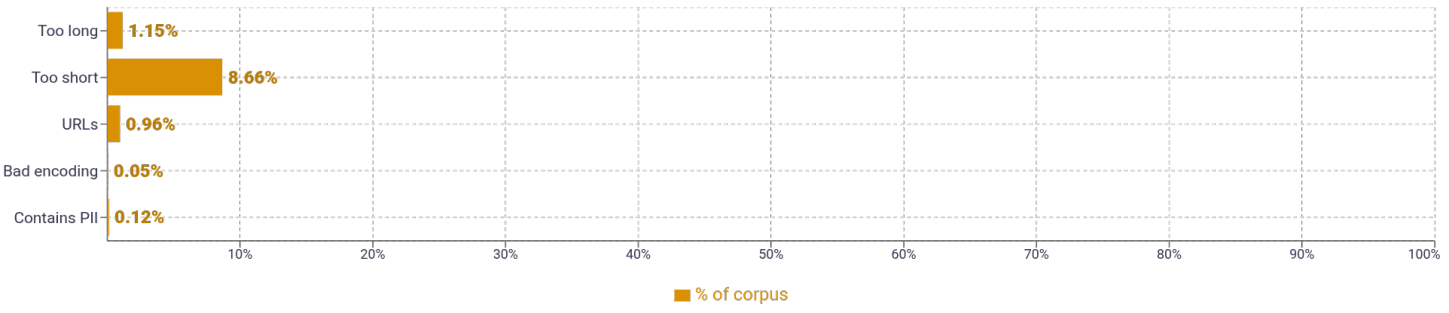
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	यांनी | 8,561,979 · करण्यात | 5,948,673 · त्यांनी | 3,829,610 · यांच्या | 3,389,817 · करत | 3,276,000
2	in marathi | 502,531 · read more | 491,831 · करू शकता | 489,138 · मोठ्या प्रमाणात | 486,006 · गुन्हा दाखल | 473,559
3	पंतप्रधान नरेंद्र मोदी | 267,190 · गुन्हा दाखल करण्यात | 192,891 · first published on | 171,657 · करण्यासाठी येथे क्लिक | 161,469 · मुख्यमंत्री उद्धव ठाकरे | 154,877
4	करण्यासाठी येथे क्लिक करा | 158,551 · जॉइन करण्यासाठी येथे क्लिक | 146,733 · ताज्या व महत्त्वाच्या बातम्या | 144,730 · क्लिक करा आणि ताज्या | 144,725 · for android and ios | 129,063
5	जॉइन करण्यासाठी येथे क्लिक करा | 145,836 · ताज्या व महत्त्वाच्या बातम्या मिळवा | 144,728 · करा आणि ताज्या व महत्त्वाच्या | 144,721 · app for android and ios | 129,038 · यूट्यूब चॅनेलला आजच सबस्क्राइब करा | 99,067

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				