HPLT Analytics report

General overview

Corpus	Analytics date	Language
mr_1.jsonl.tsv	3/25/2024	Marathi (mr)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
453,694	56,430,804	12,029,146 (21.32 %)	647M	7.56 GB	

Top 10 domains

Domain	Docs	% of total
blogspot.in	29K	6.31
transliteral.org	15K	3.26
bhaskar.com	14K	3.13
news18.com	11K	2.36
pudhari.com	10K	2.21
blogspot.com	9.4K	2.07
indiatimes.com	7.7K	1.70
majhapaper.com	6.2K	1.36
marathi.gov.in	5.8K	1.28
marathipizza.com	5.7K	1.26

Top 10 TLDs

Domain	Docs	% of total
com	293K	64.47
in	83K	18.37
org	31K	6.85
gov.in	7.6K	1.68
net	7.3K	1.62
page	6.4K	1.40
news	4.4K	0.97
co.in	4.3K	0.95
app	1.7K	0.38
info	1.5K	0.33

Documents size (in segments)

<= 25 segments **6.42%** (29K documents)
> 25 segments **93.58%** (425K documents)
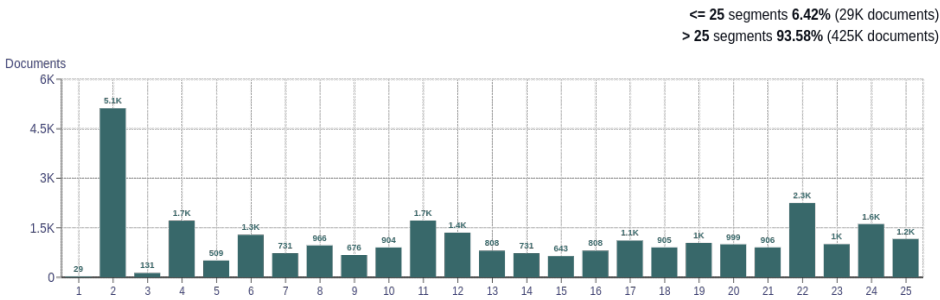
Documents by collection

cc40 (261K)
wide17 (52K)
wide16 (77K)
wide15 (64K)
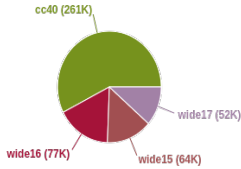
Language Distribution

Number of segments

- Marathi (mr) - 23M
- English (en) - 15M
- Hindi (hi) - 12M
- Nepali (ne) - 674K
- Czech (cs) - 624K
- Sanskrit (sa) - 609K
- Goan Konkani (gom) - 583K
- French (fr) - 504K
- Spanish (es) - 493K
- Newari (new) - 397K
- 164 Others - 2.7M
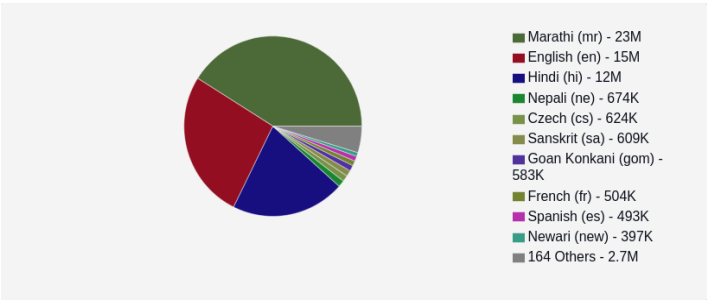
### Percentage of segments in Marathi (mr) inside documents
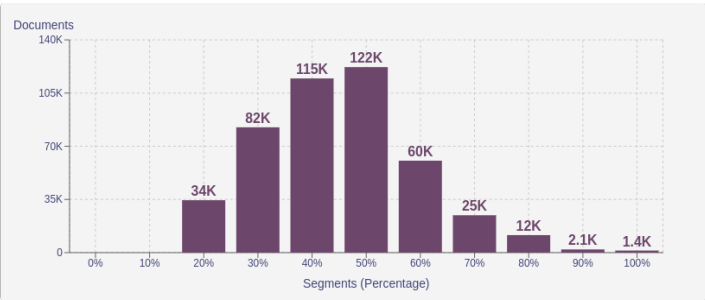
Distribution of documents by document score

score <= 5 - **24.36%** (111K documents)
score > 5 - **75.64%** (343K documents)

Segment length distribution by token

<= 49 tokens = **10M** segments | **44M** duplicates
> 50 tokens = **2.6M** segments | **649K** duplicates

Number of tokens in the segment

- Unique segments
- Duplicated segments

Segment noise distribution

Too long	0.00 %
Too short	42.45 %
URLs	1.04 %
Bad encoding	0.00 %
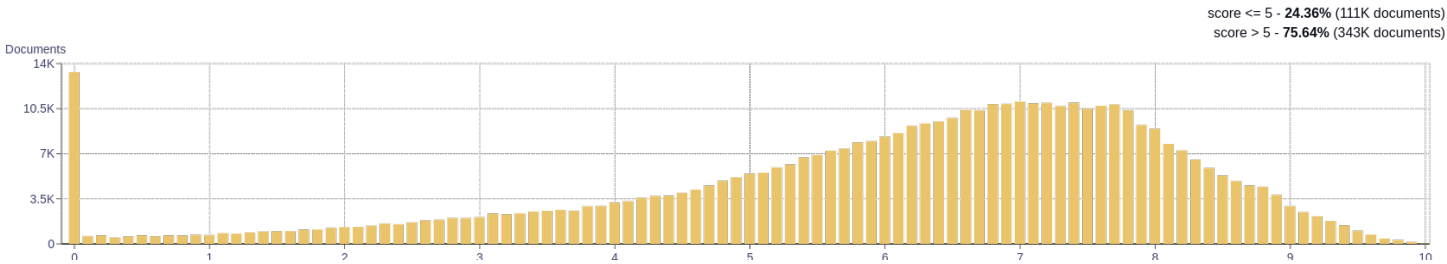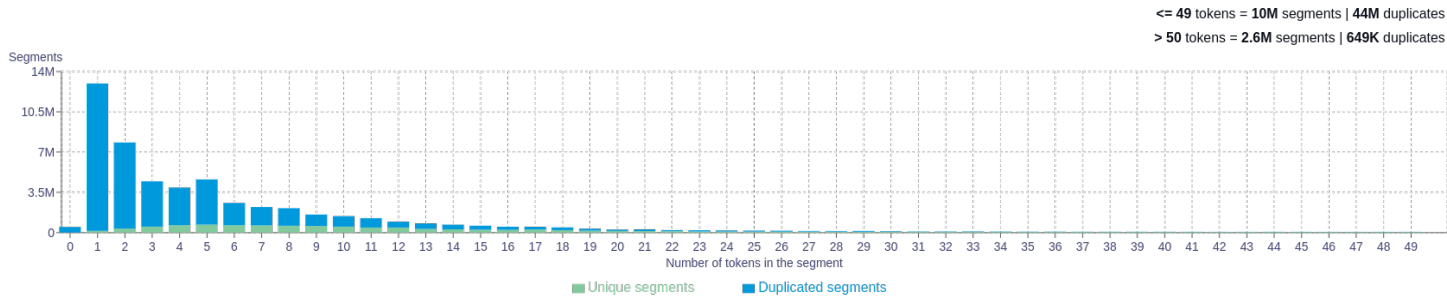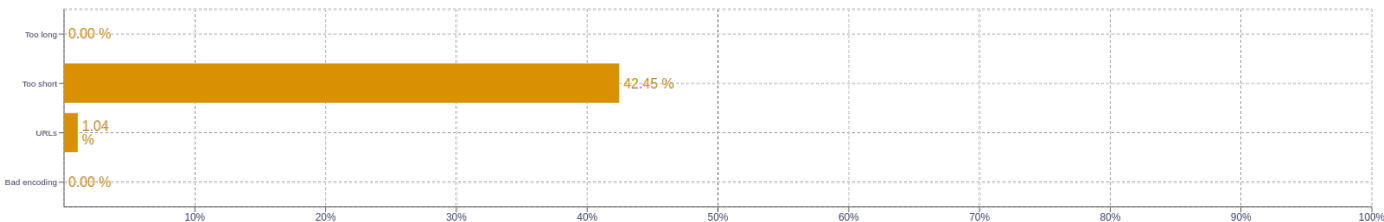
Frequent n-grams

Size	n-grams
1	to | 1598348 in | 1515505 the | 1052996 by | 987369 marathi | 980322
2	in marathi | 312738 log in | 278829 post comments | 258450 to post | 247112 or register | 238207
3	to post comments | 239262 or register to | 237950 log in or | 237830 register to post | 237727 in or register | 237316
4	or register to post | 237727 register to post comments | 237725 log in or register | 237305 in or register to | 237303 opens in new window | 118175
5	or register to post comments | 237725 log in or register to | 237303 in or register to post | 237116 to twittershare to facebookshare to | 89686 share to twittershare to facebookshare | 89686

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