HPLT Analytics report

General overview

Corpus	Analytics date	Language
mar_Deva.jsonl.tsv	9/23/2024	Marathi (mr)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
2,079,680	36,315,277	20,571,649 (56.65 %)	1.2B	16.02 GB	6,587,873,834

Top 10 domains

Domain	Docs	% of total
blogspot.in	82K	3.95
bhaskar.com	80K	3.85
esakal.com	79K	3.82
loksatta.com	67K	3.23
blogspot.com	66K	3.19
wikipedia.org	57K	2.74
news18.com	56K	2.71
indiatimes.com	37K	1.79
tv9marathi.com	33K	1.60
ibnlokmat.tv	28K	1.36

Top 10 TLDs

Domain	Docs	% of total
com	1.4M	67.81
in	330K	15.89
org	122K	5.85
tv	31K	1.49
gov.in	24K	1.16
co.in	24K	1.16
net	24K	1.13
news	21K	1.00
page	19K	0.89
es	12K	0.58

Documents size (in segments)

<= 25 segments **85.2%** (1.8M documents)
> 25 segments **14.8%** (308K documents)
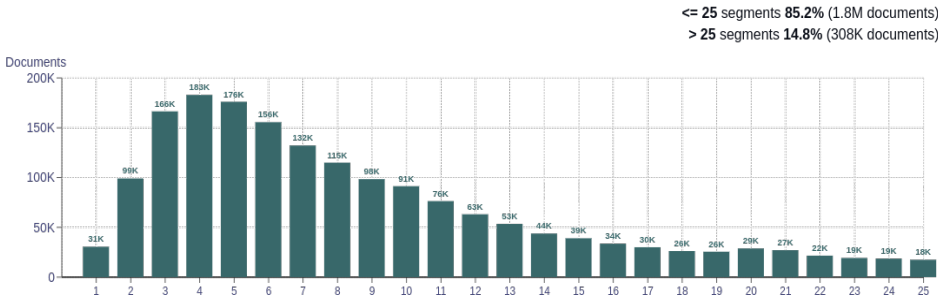
Documents by collection

cc22 (912K)
cc21 (328K)
cc18 (238K)
18 Others (602K)
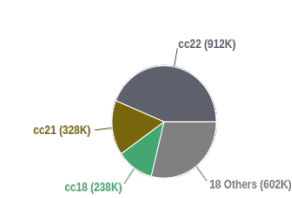
Language Distribution

Number of segments

- Marathi (mr) - 28M
- Hindi (hi) - 3.9M
- English (en) - 2.8M
- Italian (it) - 322K
- Nepali (ne) - 203K
- Goan Konkani (gom) - 172K
- Sanskrit (sa) - 161K
- French (fr) - 90K
- Newari (new) - 64K
- Portuguese (pt) - 38K
- 159 Others - 399K
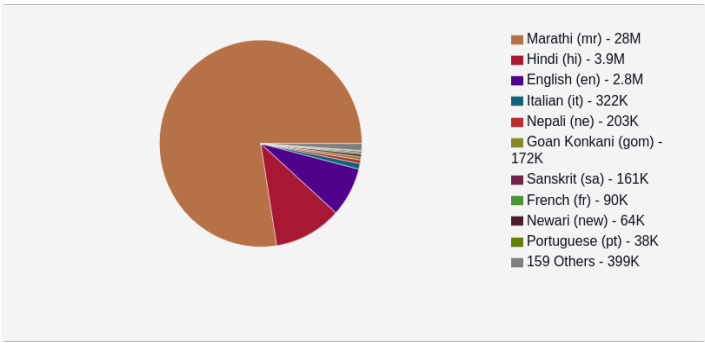
Percentage of segments in Marathi (mr) inside documents
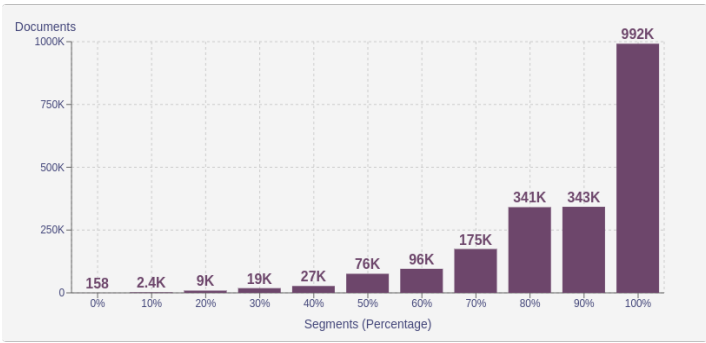
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (2.1M documents)

Segment length distribution by token

<= 49 tokens = **15M** segments | **14M** duplicates
> 50 tokens = **7.8M** segments | **2.2M** duplicates

- Unique segments
- Duplicated segments

Segment noise distribution

- Too long: 1.56 %
- Too short: 13.07 %
- URLs: 0.74 %
- Bad encoding: 0.00 %
- Contains PII: 0.12 %
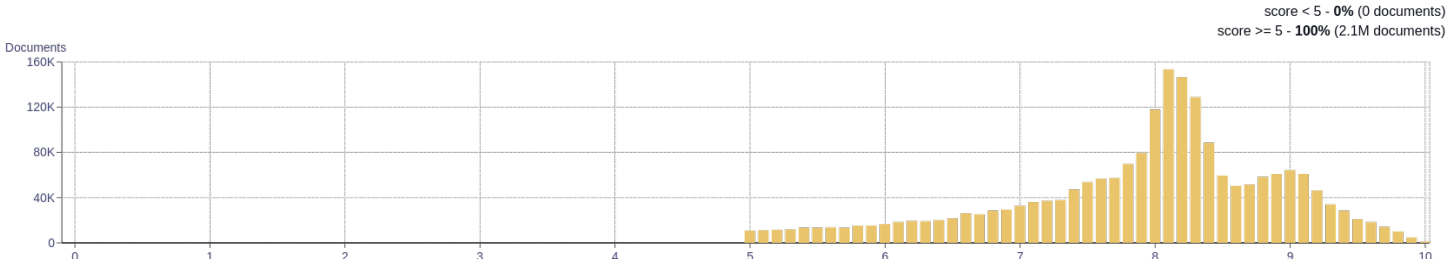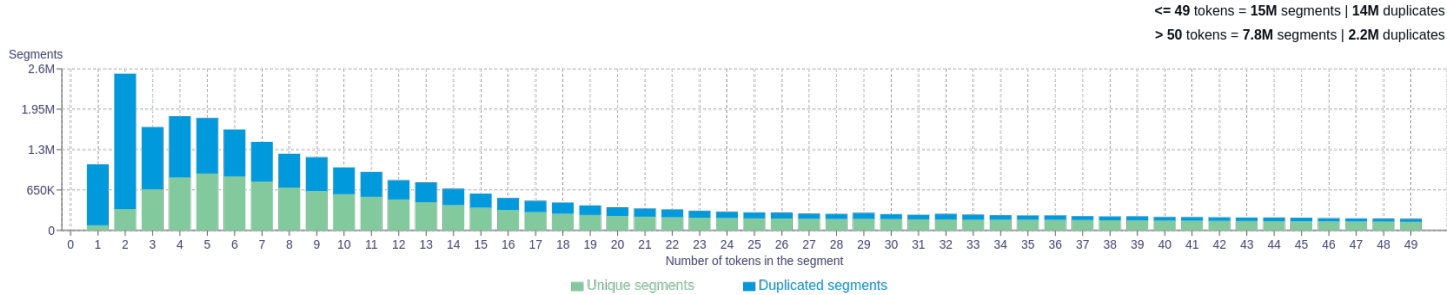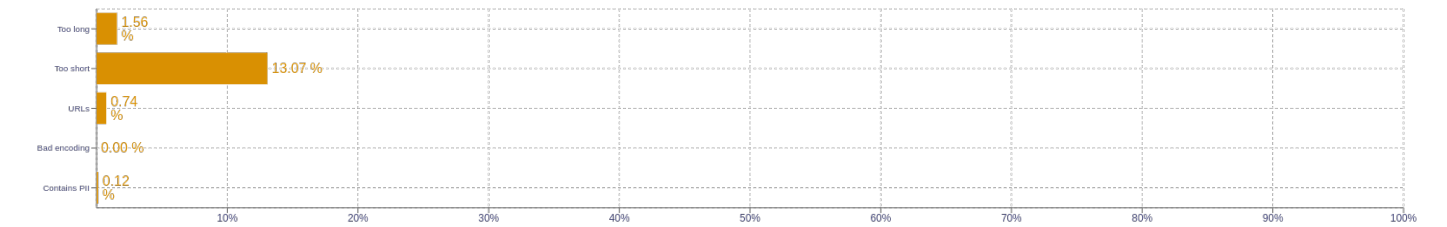
Frequent n-grams

Size	n-grams
1	यांनी | 2721206 करण्यात | 1843409 किंवा | 1820179 त्यांनी | 1588940 आपण | 1551143
2	in marathi | 213206 मोठ्या प्रमाणात | 153558 करू शकता | 134827 पुन्हा एकदा | 126837 यांनी सांगितले | 124606
3	पंतप्रधान नरेंद्र मोदी | 64747 आम्हाला फॉलो करा | 50515 all rights reserved | 48853 मुख्यमंत्री उद्धव ठाकरे | 42998 शहरातील ताज्या बातम्या | 42286
4	ताज्या बातम्या आणि ई | 42283 this website follows the | 42117 website follows the dnpa | 42115 the dnpa code of | 42113 follows the dnpa code | 42113
5	शहरातील ताज्या बातम्या आणि ई | 42283 this website follows the dnpa | 42115 website follows the dnpa code | 42113 the dnpa code of ethics | 42113 follows the dnpa code of | 42113

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