HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-v2-hin_Deva.tsv	9/23/2024	Hindi (hi)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
13,651,945	267,411,088			101.61 GB	43,703,795,598

Top 10 domains

Domain	Docs	% of total
bhaskar.com	610K	4.47
blogspot.com	435K	3.19
blogspot.in	351K	2.57
jagran.com	152K	1.11
indiatimes.com	133K	0.97
alibaba.com	128K	0.94
amarujala.com	120K	0.88
punjabkesari.in	110K	0.81
wikipedia.org	95K	0.70
webdunia.com	91K	0.67

Top 10 TLDs

Domain	Docs	% of total
com	9.6M	70.08
in	2.3M	16.61
org	581K	4.26
page	233K	1.71
co.in	189K	1.38
net	156K	1.14
news	53K	0.39
info	47K	0.34
co	46K	0.34
ae	39K	0.28

Documents size (in segments)

<= 25 segments **81.24%** (11M documents)
> 25 segments **18.76%** (2.6M documents)
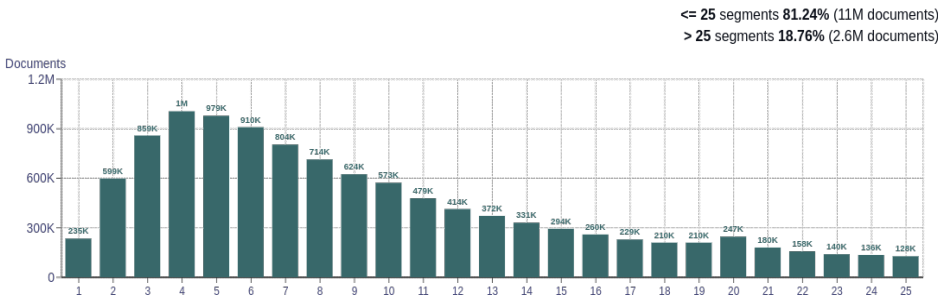
Documents by collection

cc22 (5.4M)
cc18 (1.9M)
cc21 (1.9M)
18 Others (4.4M)
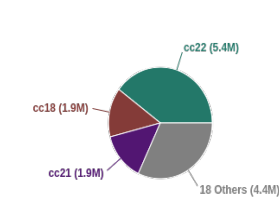
Language Distribution

Number of segments

- Hindi (hi) - 231M
- English (en) - 19M
- Nepali (ne) - 5.8M
- Italian (it) - 3.6M
- Marathi (mr) - 1.1M
- Sanskrit (sa) - 754K
- French (fr) - 750K
- German (de) - 585K
- Newari (new) - 499K
- Spanish (es) - 294K
- 165 Others - 3.5M
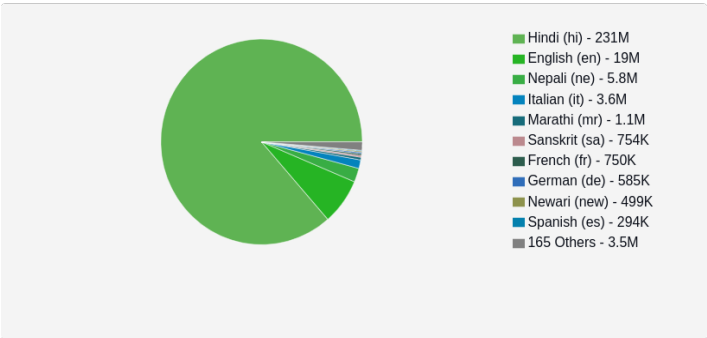
Percentage of segments in Hindi (hi) inside documents
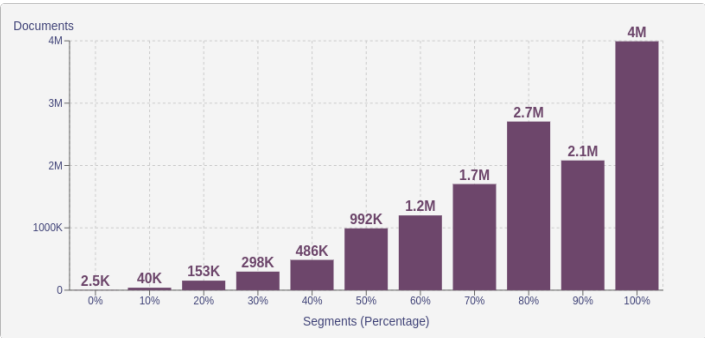
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (14M documents)

Segment noise distribution

Too long	0.00 %
Too short	7.02 %
URLs	0.76 %
Bad encoding	0.00 %
Contains PII	0.11 %
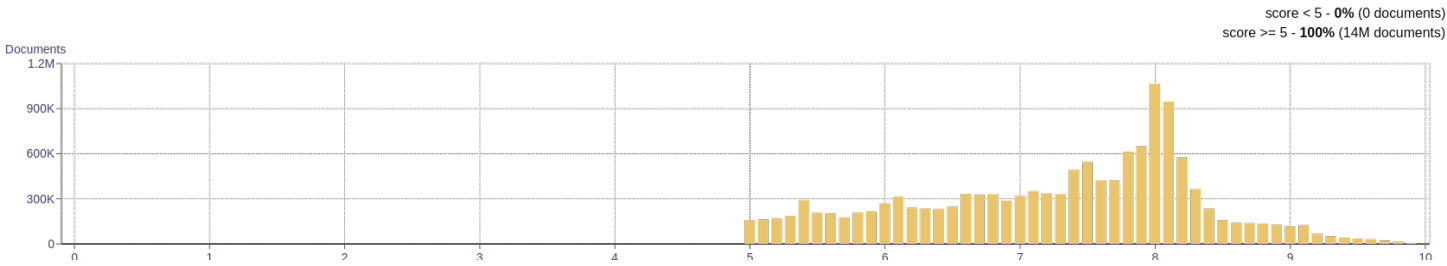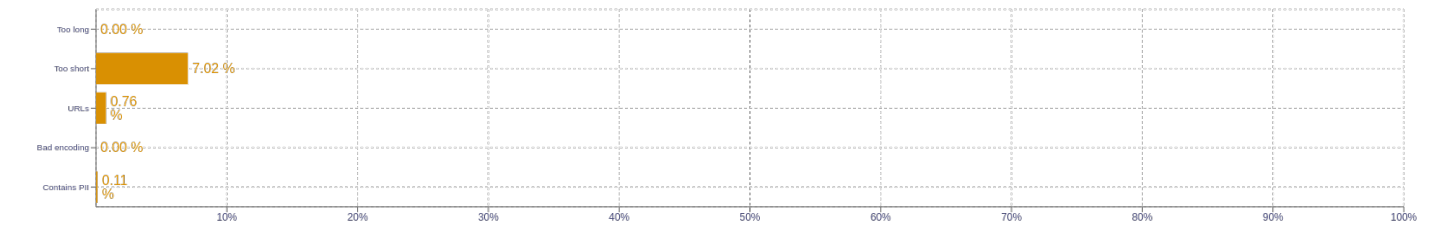
About HPLT Analytics

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