HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-v2-ben_Beng.tsv	9/19/2024	Bangla (bn)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
11,043,918	176,013,069			73.69 GB	29,990,290,840

Top 10 domains

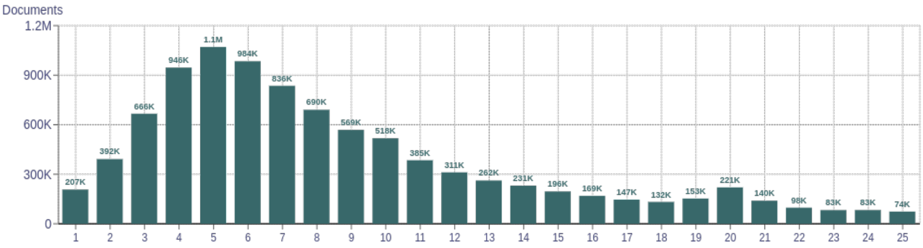
Domain	Docs	% of total
wikipedia.org	137K	1.24
bdnews24.com	93K	0.84
kalerkantho.com	85K	0.77
dailyjanakantha.com	79K	0.72
anandabazar.com	73K	0.66
banglanews24.com	65K	0.59
deshebideshe.com	61K	0.56
ournewsbd.com	57K	0.51
news18.com	57K	0.51
blogspot.com	54K	0.49

Top 10 TLDs

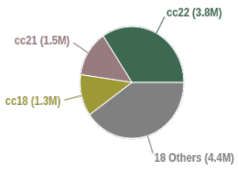
Domain	Docs	% of total
com	8.9M	80.90
net	514K	4.65
org	384K	3.48
in	257K	2.33
com.bd	232K	2.10
tv	133K	1.20
news	105K	0.95
gov.bd	72K	0.65
info	42K	0.38
ru	22K	0.20

Documents size (in segments)

<= 25 segments **86.58%** (9.6M documents)
> 25 segments **13.42%** (1.5M documents)

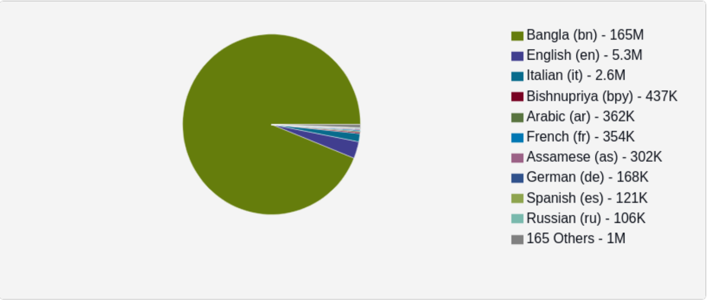


Documents by collection

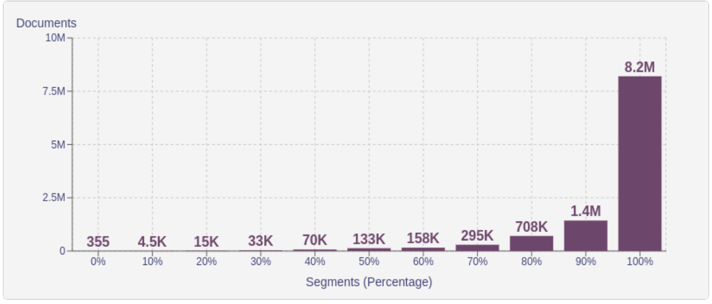


Language Distribution

Number of segments

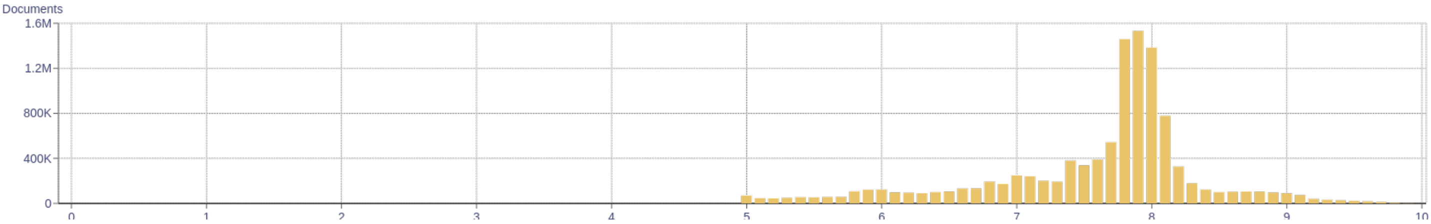


Percentage of segments in Bangla (bn) inside documents

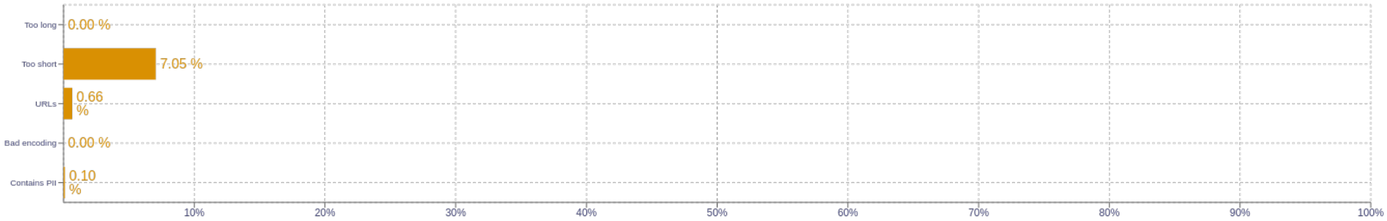


Distribution of documents by document score

score <= 5 - **100%** (11M documents)
score > 5 - **0%** (509 documents)



Segment noise distribution



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