

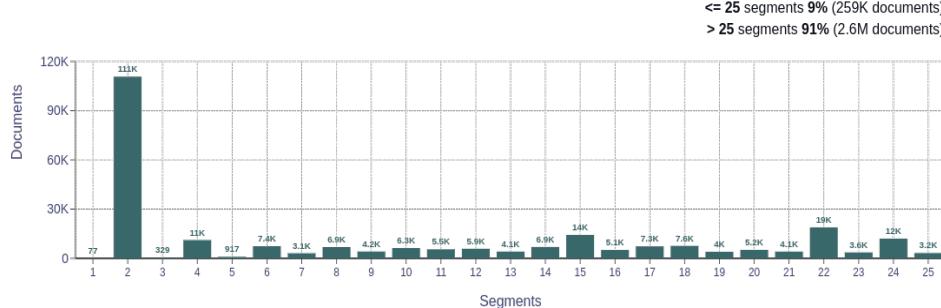
General overview

Corpus	Analytics date	Language
HPLT-docsite.bn.tsv	6/8/2024	Bangla (bn)

Volumes

Docs	Segments	Unique segments	Tokens	Size
2,875,658	330,215,118	79,924 (0.02 %)	3.2B	40.7 GB

Documents size (in segments)



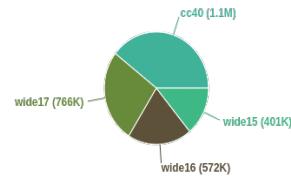
Top 10 domains

Domain	Docs	% of total
chanood.com	30K	1.03
fanpop.com	22K	0.76
anandabazar.com	19K	0.66
blogspot.in	17K	0.58
blogspot.com	16K	0.55
wikipedia.org	15K	0.53
newspapers71.com	15K	0.52
sangbadpratidin.in	14K	0.50
bn-takedrivers.com	13K	0.47
bdnews24.com	12K	0.42

Top 10 TLDs

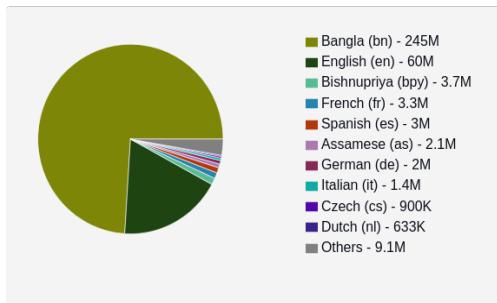
Domain	Docs	% of total
.com	2.4M	83.06
.net	95K	3.29
.in	80K	2.77
.org	76K	2.64
.com.bd	49K	1.69
.tv	23K	0.80
.news	19K	0.64
.info	11K	0.39
.xyz	9.5K	0.33
.top	8.9K	0.31

Documents by collection

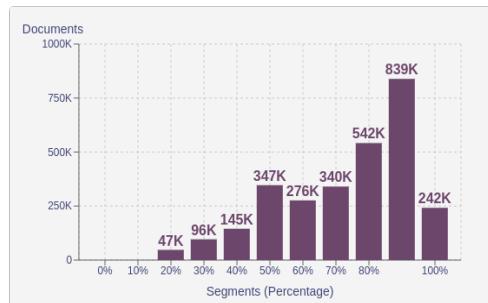


Language Distribution

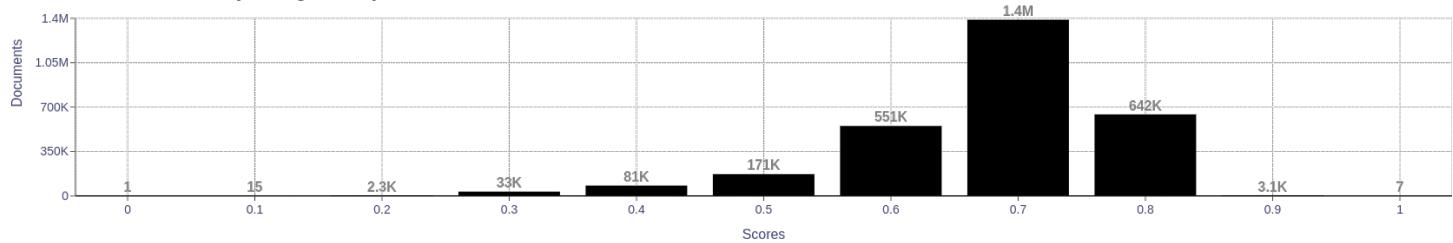
Number of segments



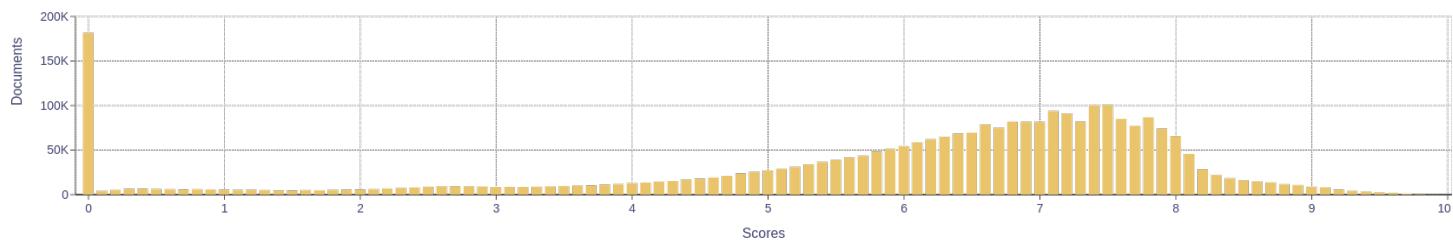
Percentage of segments in Bangla (bn) inside documents



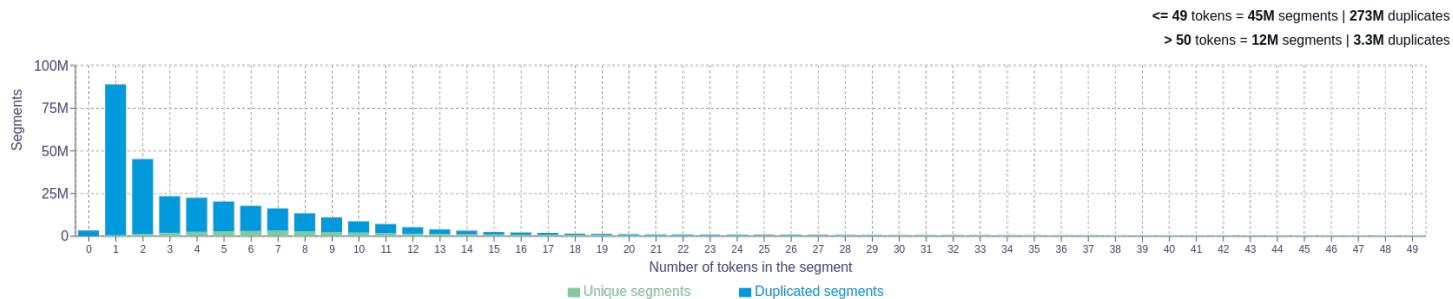
Distribution of documents by average fluency score



Distribution of documents by document score



Segment length distribution by token



About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , etc.) replaced by newlines.

Language distribution

Language identified with FastSpell (<https://github.com/mabanon/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>