HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
ben_Beng.jsonl.tsv	6/11/2025	Bangla (bn)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
11,043,918	175,970,018	98,288,754 (55.86 %)	5.3B	29,990,290,840	73.69 GB

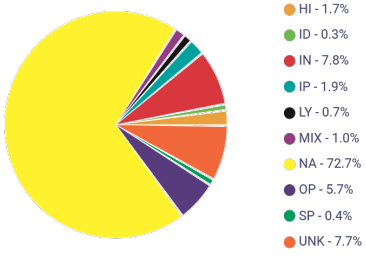
Top 10 domains

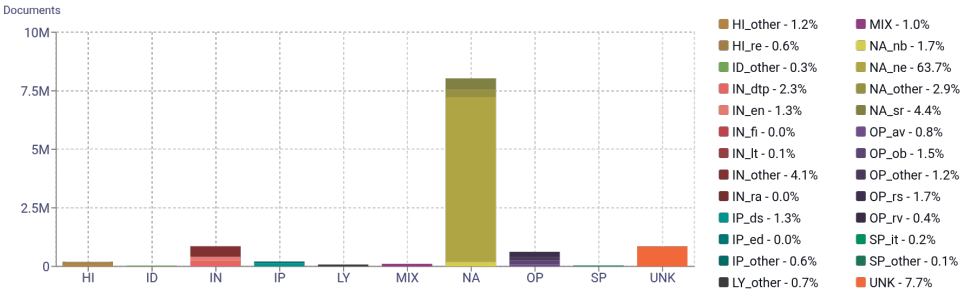
Domain	Docs	% of total
wikipedia.org	137K	1.24%
bdnews24.com	93K	0.84%
kalerkantho.com	85K	0.77%
dailyjanakantha...	79K	0.72%
anandabazar.com	73K	0.66%
banglanews24.com	65K	0.59%
deshebideshe.com	61K	0.56%
ournewsbd.com	57K	0.51%
news18.com	57K	0.51%
blogspot.com	54K	0.49%

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%

Register labels

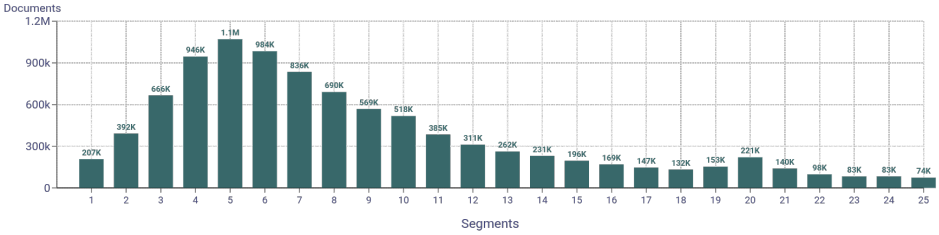




MT:3.1% | 343K Documents

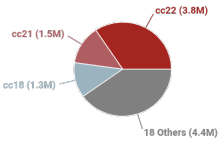
Documents size (in segments)

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



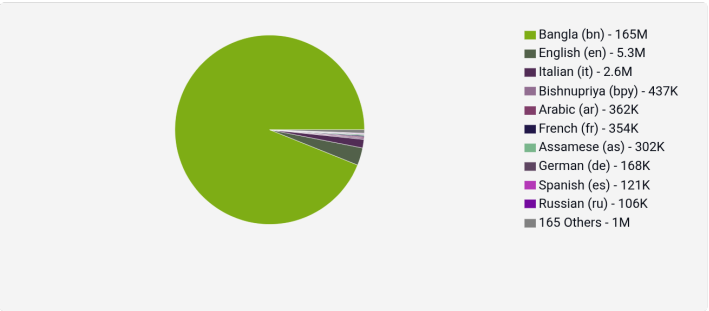
Documents by collection

CC = 66.29%
IA = 33.71%

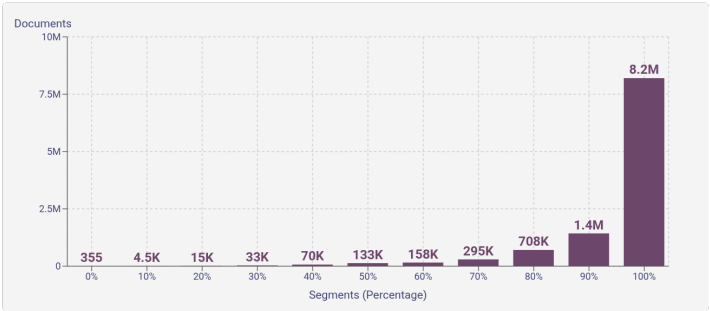


Language Distribution

Number of segments in the Bangla (bn) corpus

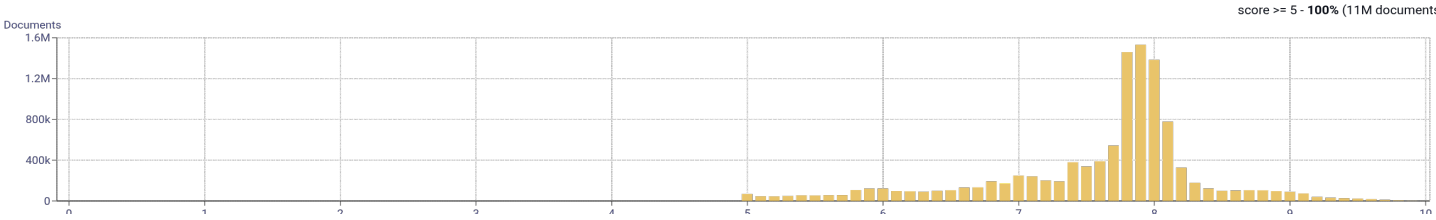


Percentage of segments in Bangla (bn) inside documents



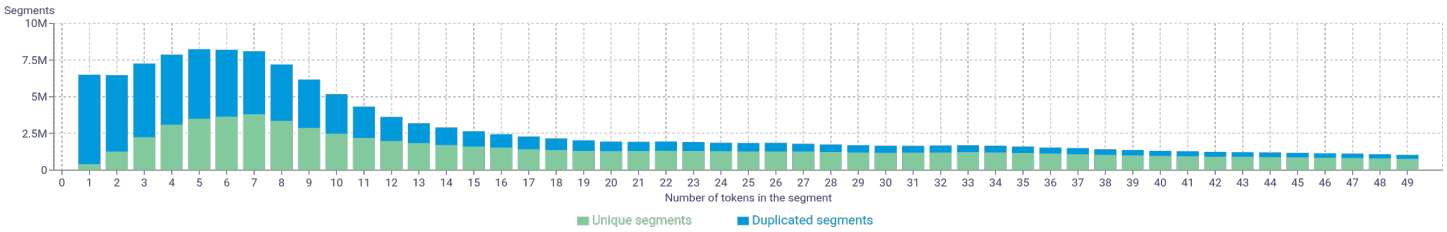
Distribution of documents by document score

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

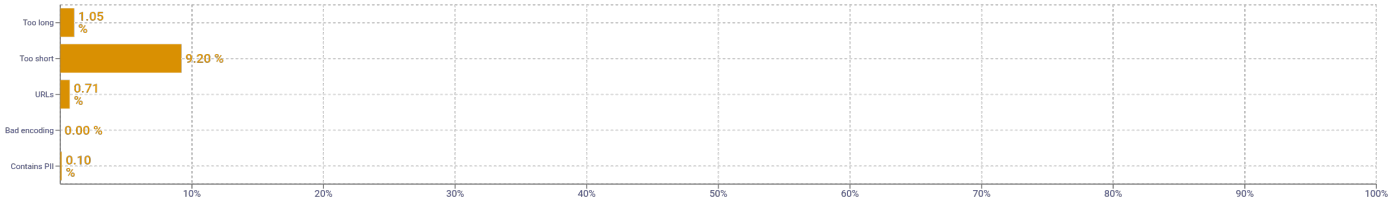


Segment length distribution by token

≤ 49 tokens = **74M** segments | **69M** duplicates
> 50 tokens = **33M** segments | **9.1M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	এক | 13338288 · সময় | 9365016 · কথা | 8755173 · সাথে | 8347240 · একটা | 5869494
2	আওয়ামী লীগের | 1576059 · আওয়ামী লীগ | 1010892 · শেখ হাসিনা | 993009 · প্রধানমন্ত্রী শেখ | 860632 · read more | 606496
3	প্রধানমন্ত্রী শেখ হাসিনা | 532443 · বঙ্গবন্ধু শেখ মুজিবুর | 352490 · প্রধানমন্ত্রী শেখ হাসিনার | 259445 · অতিথি হিসেবে উপস্থিত | 249104 · থানার ভারপ্রাপ্ত কর্মকর্তা | 243289
4	আওয়ামী লীগের সাধারণ সম্পাদক | 203503 · বঙ্গবন্ধু শেখ মুজিবুর রহমানের | 189018 · প্রধান অতিথি হিসেবে উপস্থিত | 151749 · বঙ্গবন্ধু শেখ মুজিবুর রহমান | 129697 · জাতির পিতা বঙ্গবন্ধু শেখ | 123804
5	জাতির পিতা বঙ্গবন্ধু শেখ মুজিবুর | 119497 · মহাসচিব মির্জা ফখরুল ইসলাম আলমগীর | 89229 · জাতির জনক বঙ্গবন্ধু শেখ মুজিবুর | 84052 · পিতা বঙ্গবন্ধু শেখ মুজিবুর রহমানের | 75469 · জনক বঙ্গবন্ধু শেখ মুজিবুর রহমানের | 54963

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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av