

General overview

Corpus	Date	Language
asm_Beng.jsonl.tsv	9/7/2024	Assamese (as)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
175,709	2,676,868	1,762,630 (65.85 %)	87M	473,157,298	1.15 GB

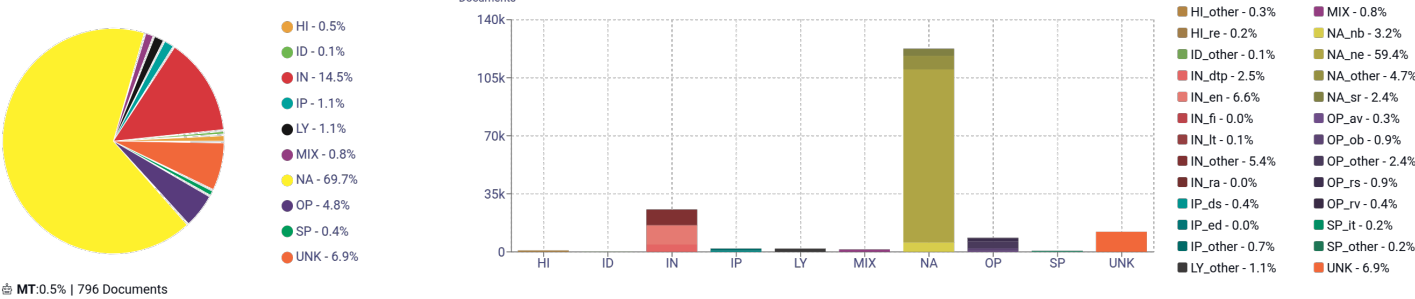
Top 10 domains

Domain	Docs	% of total
nenow.in	28K	15.71%
news18.com	16K	8.85%
eastmojo.com	12K	6.95%
wikipedia.org	12K	6.63%
sentinelassam.com	8.1K	4.61%
xahitya.org	8K	4.54%
xukhdukh.com	7K	3.98%
janambhumi.in	4.5K	2.56%
nefocus.com	4.1K	2.35%
dainikagradoot.in	3.6K	2.03%

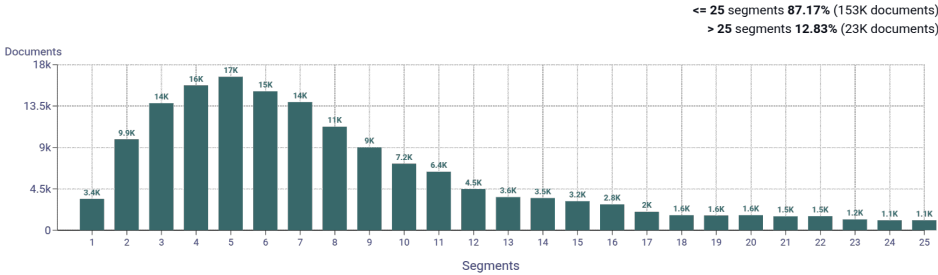
Top 10 TLDs

Domain	Docs	% of total
com	93K	53.19%
in	55K	31.12%
org	24K	13.44%
co.in	1.1K	0.61%
gov.in	1K	0.59%
org.in	463	0.26%
info	267	0.15%
online	222	0.13%
net	201	0.11%
ae	181	0.10%

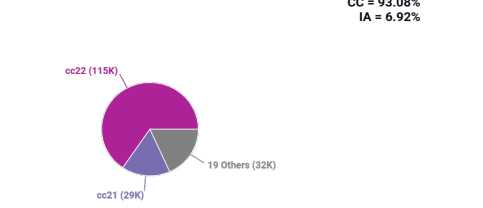
Register labels



Documents size (in segments)

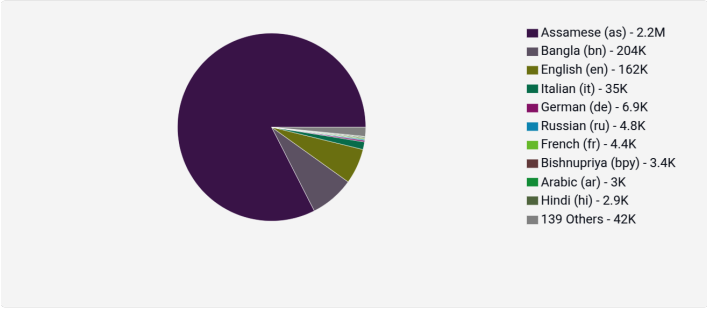


Documents by collection

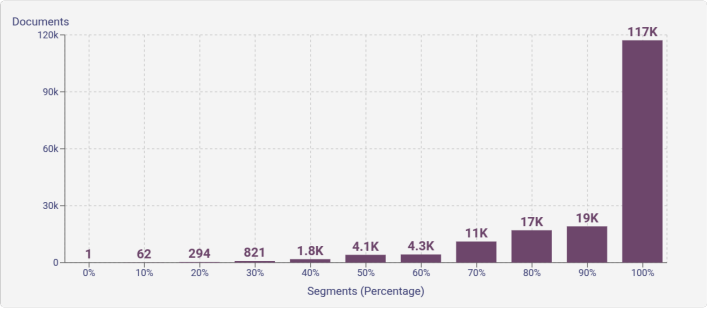


Language Distribution

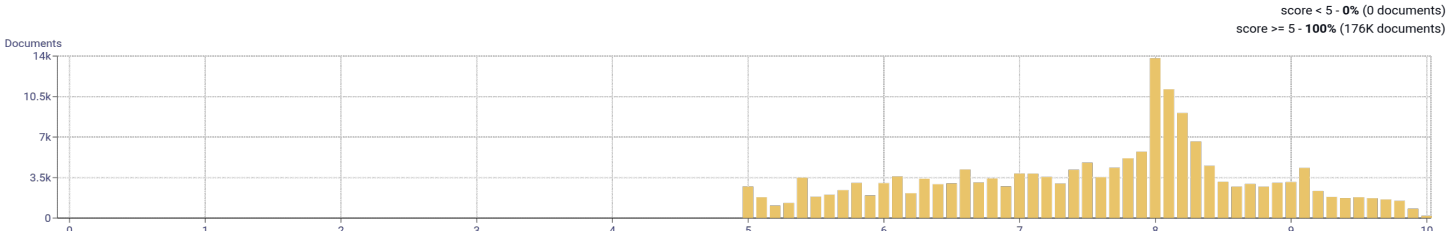
Number of segments in the Assamese (as) corpus



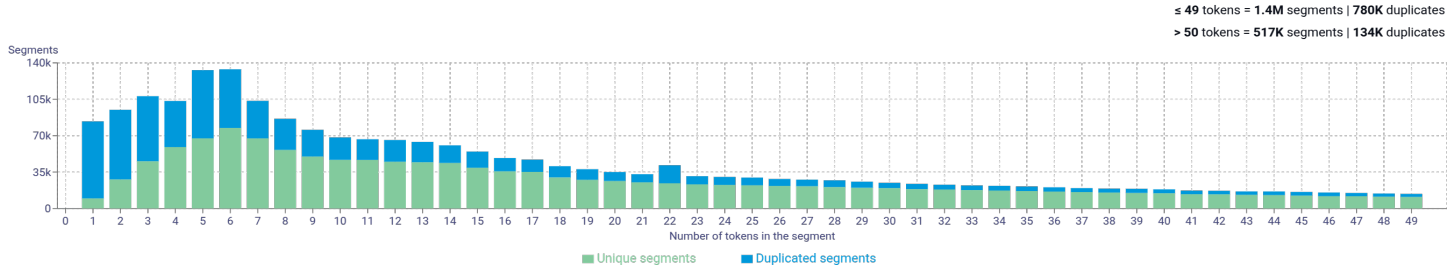
Percentage of segments in Assamese (as) inside documents



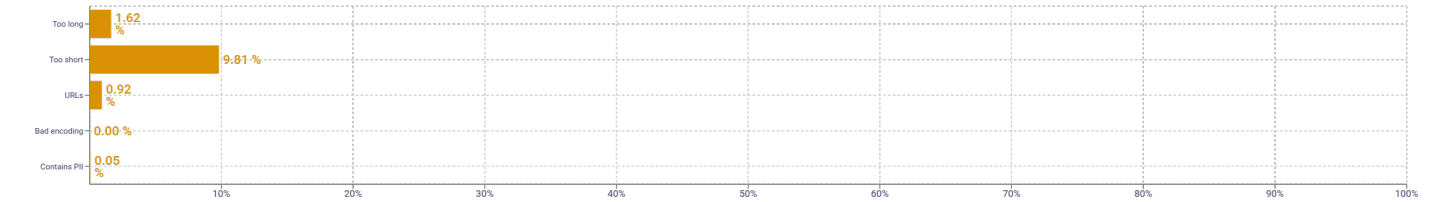
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	এই 600173 ক'ৰা 514063 ক'ৰি 463001 হ 430200 হা 369615
2	ক'ৰা হা 76763 ক'ৰা হৈছে 71132 সম্পাদনা কৰক 57284 ক'ৰিব পাৰে 35000 ক'ৰা হ 31376
3	জাতিৰ পৰা গৈছে 11483 লাইভ নিউজ আপডেট 8941 সবাতোকৈ বিশ্বাসযোগ্য অসমীয়া 8940 শ্বেকিং নিউজ সন্সখন 8940 বিশ্বাসযোগ্য অসমীয়া নিউজ 8940
4	সবাতোকৈ বিশ্বাসযোগ্য অসমীয়া নিউজ 8940 শ্বেকিং নিউজ সন্সখন news18 8940 বিশ্বাসযোগ্য অসমীয়া নিউজ বেছাইট 8940 অসমীয়াত শ্বেকিং নিউজ সন্সখন 8940 অসমীয়া নিউজ বেছাইট news18 8940
5	সবাতোকৈ বিশ্বাসযোগ্য অসমীয়া নিউজ বেছাইট 8940 বিশ্বাসযোগ্য অসমীয়া নিউজ বেছাইট news18 8940 অসমীয়াত শ্বেকিং নিউজ সন্সখন news18 8940 শ্বেকিং নিউজ সন্সখন news18 অসমীয়াত 8891 অসমীয়া নিউজ বেছাইট news18 অসমীয়া 8891

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 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>, , , 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/pablopt6n/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.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	How-to or instructions	HI	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	Informational persuasion	IP	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	Opinion	OP
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	Informational description	IN	Opinion blog	ob
News report	ne	Encyclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				