HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

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

Top 10 domains

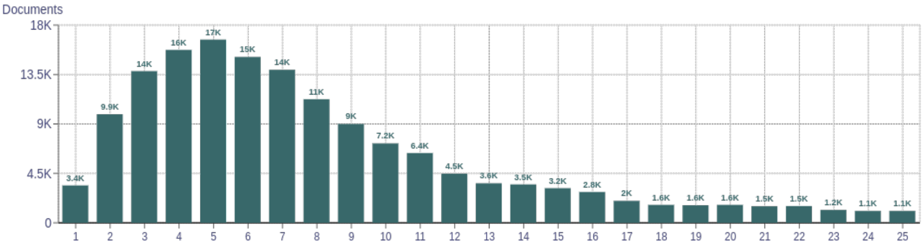
Domain	Docs	% of total
nenow.in	28K	15.71
news18.com	16K	8.85
eastmojo.com	12K	6.95
wikipedia.org	12K	6.63
sentinelassam.com	8.1K	4.61
xahitya.org	8K	4.54
xukhdukh.com	7K	3.98
janambhumi.in	4.5K	2.56
nefocus.com	4.1K	2.35
dainikagradoot.in	3.6K	2.03

Top 10 TLDs

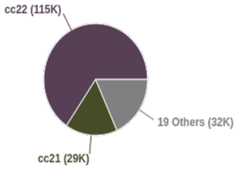
Domain	Docs	% of total
com	93K	53.19
in	55K	31.12
org	24K	13.44
co.in	1.1K	0.61
gov.in	1K	0.59
org.in	463	0.26
info	267	0.15
online	222	0.13
net	201	0.11
ae	181	0.10

Documents size (in segments)

<= 25 segments **87.17%** (153K documents)
> 25 segments **12.83%** (23K documents)

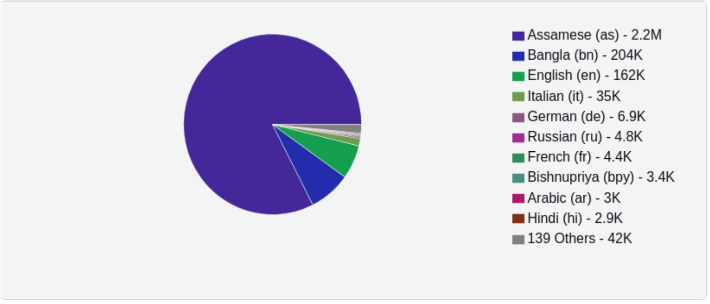


Documents by collection

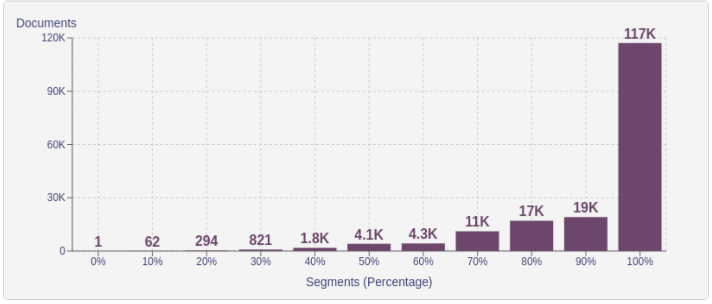


Language Distribution

Number of segments

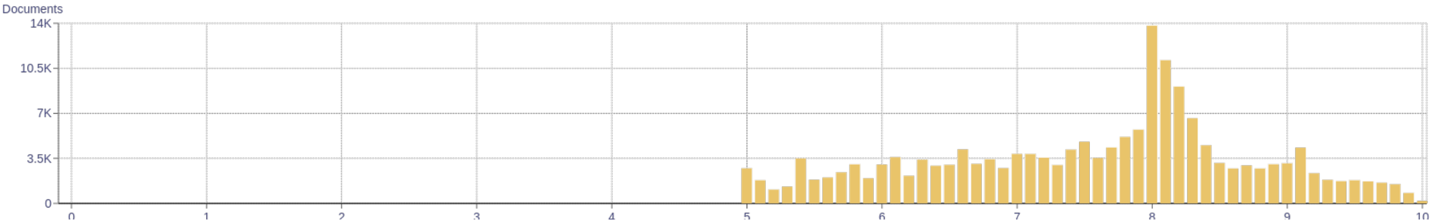


Percentage of segments in Assamese (as) inside documents



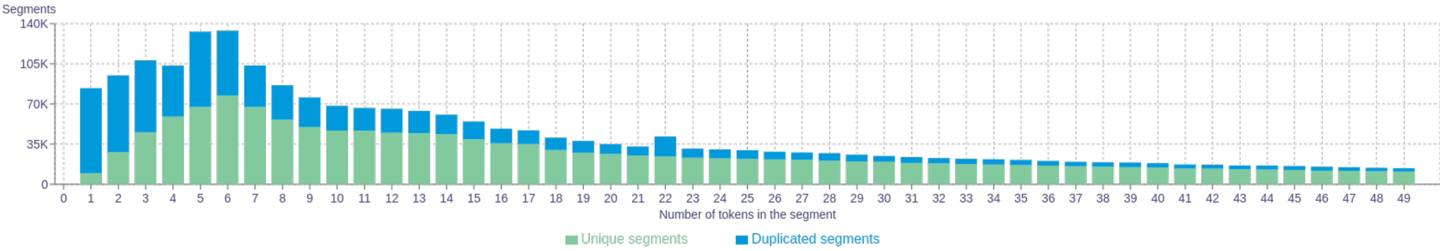
Distribution of documents by document score

score <= 5 - **99.89%** (176K documents)
score > 5 - **0.11%** (195 documents)

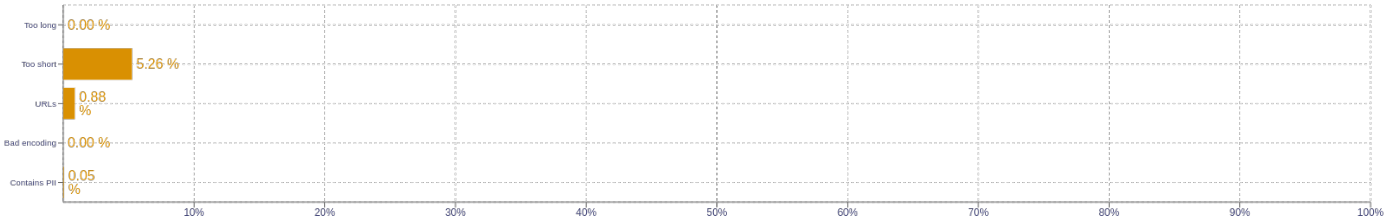


Segment length distribution by token

<= 49 tokens = **1.4M** segments | **780K** duplicates
> 50 tokens = **517K** segments | **134K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	এই | 600173 কৰা | 514063 কৰি | 463001 হ | 430200 হয় | 369615
2	কৰা হয় | 76763 কৰা হৈছে | 71132 সম্পাদনা কৰক | 57284 কৰিব পাৰে | 35000 কৰা হ | 31376
3	জানিব পৰা গৈছে | 11483 লাইভ নিউজ আপডেট | 8941 সবাতোকৈ বিশ্বাসযোগ্য অসমীয়া | 8940 ব্ৰেকিং নিউজ সৰ্বপ্ৰথম | 8940 বিশ্বাসযোগ্য অসমীয়া নিউজ | 8940
4	সবাতোকৈ বিশ্বাসযোগ্য অসমীয়া নিউজ | 8940 ব্ৰেকিং নিউজ সৰ্বপ্ৰথম news18 | 8940 বিশ্বাসযোগ্য অসমীয়া নিউজ ৱেবছাইট | 8940 অসমীয়াত ব্ৰেকিং নিউজ সৰ্বপ্ৰথম | 8940 অসমীয়া নিউজ ৱেবছাইট news18 | 8940
5	সবাতোকৈ বিশ্বাসযোগ্য অসমীয়া নিউজ ৱেবছাইট | 8940 বিশ্বাসযোগ্য অসমীয়া নিউজ ৱেবছাইট news18 | 8940 অসমীয়াত ব্ৰেকিং নিউজ সৰ্বপ্ৰথম news18 | 8940 ব্ৰেকিং নিউজ সৰ্বপ্ৰথম news18 অসমীয়াত | 8891 অসমীয়া নিউজ ৱেবছাইট news18 অসমীয়া | 8891

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