

General overview

Corpus	Analytics date	Language
mni_Beng.jsonl.tsv	12/6/2024	Manipuri (mni)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
2,934	65,762	40,126 (61.02 %)	1.9M	29.36 MB	11,728,425

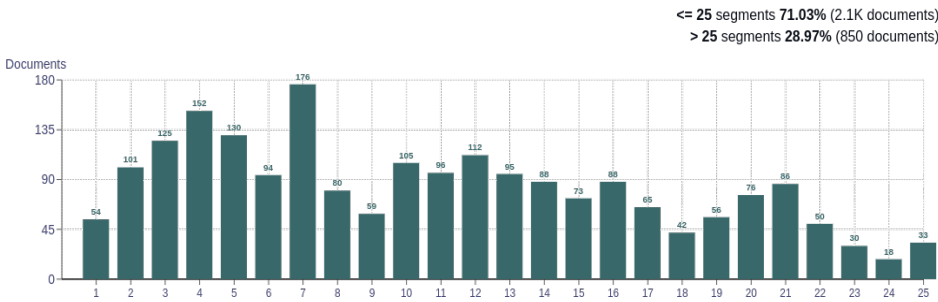
Top 10 domains

Domain	Docs	% of total
pib.gov.in	748	25.49
narendramodi.in	642	21.88
lakhipuronline.in	535	18.23
vikaspedia.in	348	11.86
manipurimirror.com	151	5.15
naharolghithoudang.in	69	2.35
jw.org	65	2.22
gotquestions.org	53	1.81
bahai.in	40	1.36
pmindia.gov.in	33	1.12

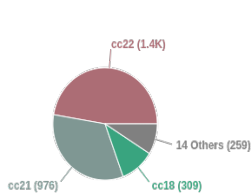
Top 10 TLDs

Domain	Docs	% of total
in	1.6K	55.79
gov.in	781	26.62
nic.in	199	6.78
com	187	6.37
org	126	4.29
net	2	0.07
gov.bd	1	0.03
ভারত	1	0.03

Documents size (in segments)

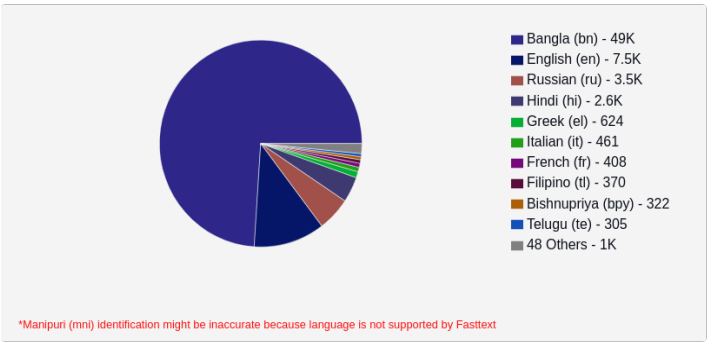


Documents by collection

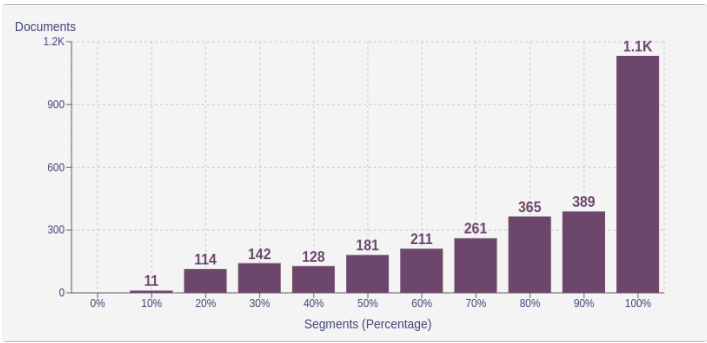


Language Distribution

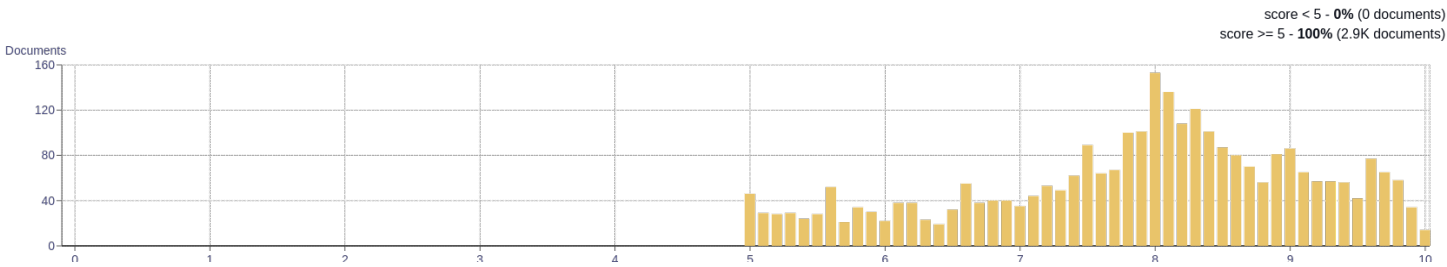
Number of segments



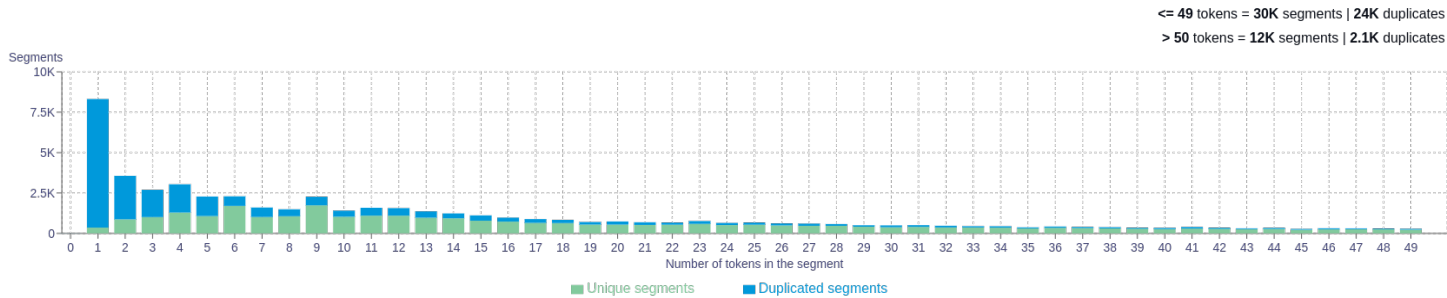
Percentage of segments in Manipuri (mni) inside documents



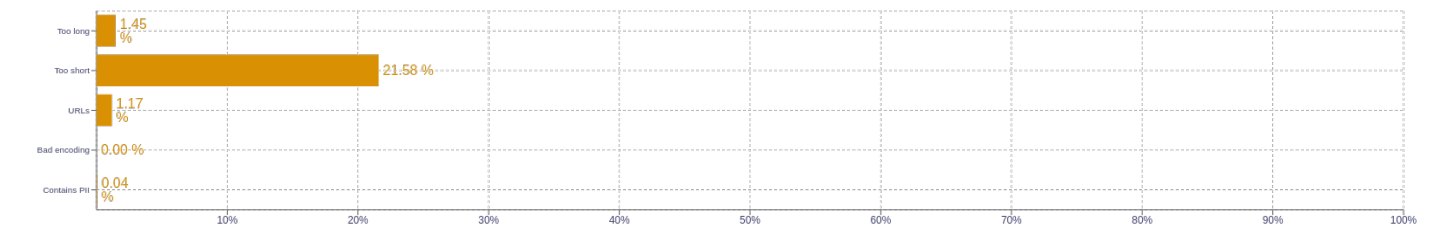
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>অমসুং 40811</div> <div>অসি 20441</div> <div>অমা 12021</div> <div>অসিনা 8575</div> <div>অসিনা 8013</div>
2	<div>প্রধান মন্ত্রীনা 2777</div> <div>লৈবাক অসিগী 1652</div> <div>pmo india 1496</div> <div>লৈবাক অসিনা 1091</div> <div>মৰু ওইবা 1088</div>
3	<div>by pib imphal 748</div> <div>প্রধান মন্ত্রীনা হায়াথি 559</div> <div>শ্রী নরেন্দ্র মোদীনা 558</div> <div>হাঘনা প্রধান মন্ত্রীনা 528</div> <div>প্রধান মন্ত্রী শ্রী 443</div>
4	<div>প্রধান মন্ত্রী শ্রী নরেন্দ্র মোদীনা 404</div> <div>read this release in 306</div> <div>বরাক তম্পাঞ্জী খোমজিনবা পাউ 169</div> <div>প্রধান মন্ত্রীনা হায়াথি মদুদি 167</div> <div>হকশেল অমসুং রাইফ-মুখাল মড্ডোলয় 154</div>
5	<div>প্রধান মন্ত্রী শ্রী নরেন্দ্র মোদীনা 329</div> <div>হৌথিবা পুং ২৪দা অদৌবা কেস 75</div> <div>asi ga loinana atei e-pao 74</div> <div>sing asi ga loinana atei 72</div> <div>pao sing asi ga loinana 72</div>

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.stinfo.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/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>