HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
mni_Beng.jsonl.tsv	11/7/2024	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

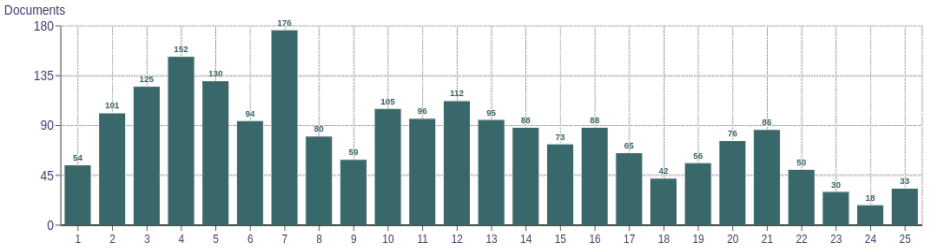
Domain	Docs	% of total
pib.gov.in	748	25.49
narendramodi.in	642	21.88
lakhipuronline.in	535	18.23
vikaspedia.in	348	11.86
manipurimirror.com	151	5.15
naharolgithoudang.in	69	2.35
jw.org	65	2.22
gotquestions.org	53	1.81
bahai.in	40	1.36
pmindia.gov.in	33	1.12

Top 10 TLDs

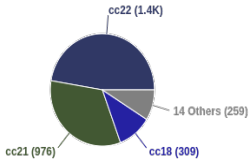
Domain	Docs	% of total
in	1.6K	55.79
gov.in	781	26.62
nic.in	199	6.78
com	187	6.37
org	126	4.29
net	2	0.07
gov.bd	1	0.03
ভারত	1	0.03

Documents size (in segments)

<= 25 segments **71.03%** (2.1K documents)
> 25 segments **28.97%** (848 documents)

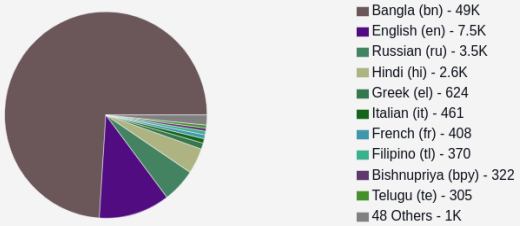


Documents by collection



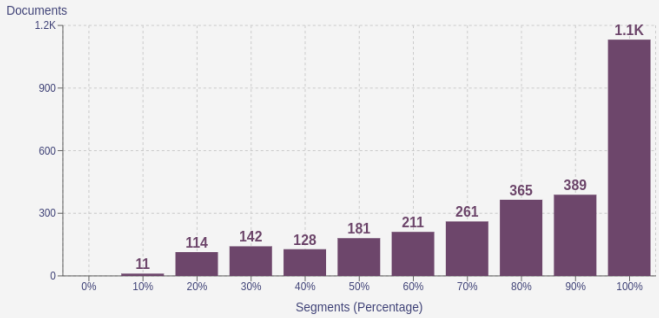
Language Distribution

Number of segments



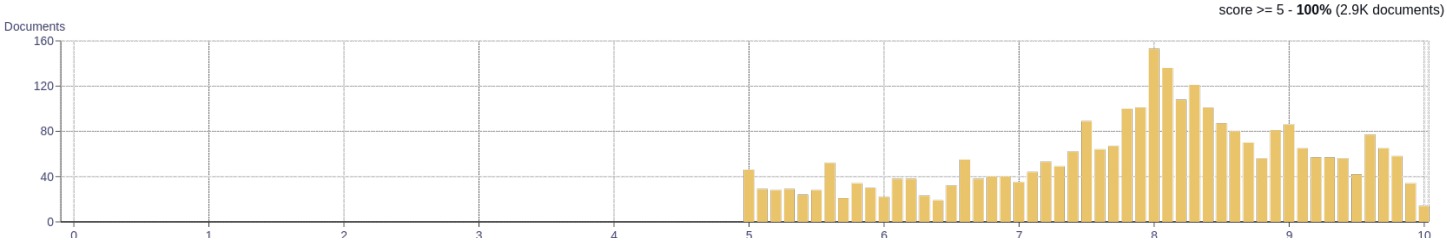
*Manipuri (mni) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Manipuri (mni) inside documents



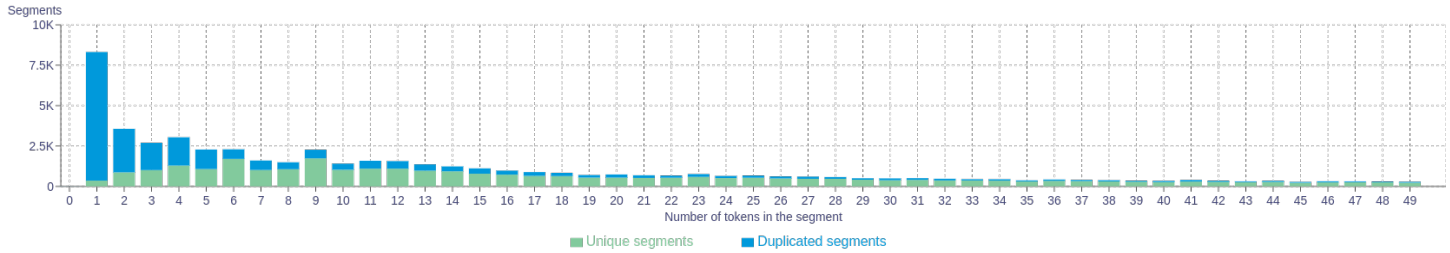
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (2.9K documents)

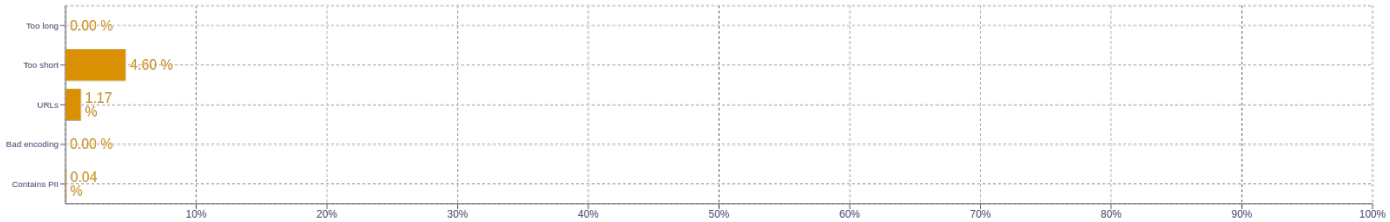


Segment length distribution by token

<= 49 tokens = **30K** segments | **24K** duplicates
> 50 tokens = **12K** segments | **2.1K** duplicates



Segment noise distribution



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