

General overview

Corpus	Analytics date	Language
hne_Deva.jsonl.tsv	11/27/2024	Chhattisgarhi (hne)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
2,806	54,999	40,771 (74.13 %)	2.5M	24.6 MB	10,541,012

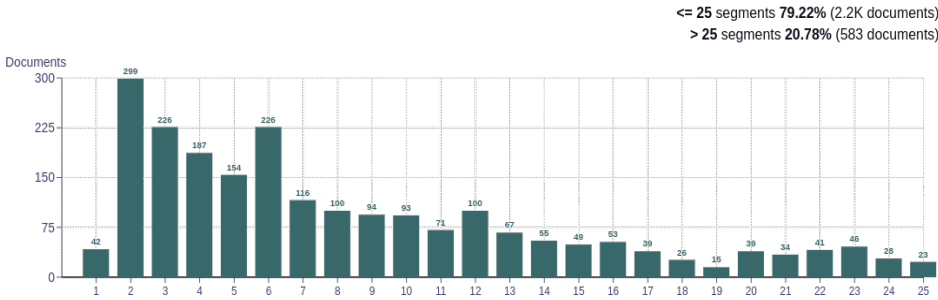
Top 10 domains

Domain	Docs	% of total
jayjohar.com	560	19.96
gurturgoth.com	462	16.46
blogspot.com	313	11.15
hi-takdrivers.com	179	6.38
blogspot.in	120	4.28
ruralindiaonline.org	117	4.17
news18.com	100	3.56
vithivichar.com	90	3.21
biblica.com	82	2.92
silverweed.club	67	2.39

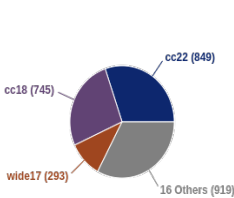
Top 10 TLDs

Domain	Docs	% of total
com	2.1K	75.37
in	185	6.59
org	169	6.02
club	159	5.67
co.in	64	2.28
online	55	1.96
ae	17	0.61
xyz	15	0.53
net	12	0.43
jp	2	0.07

Documents size (in segments)

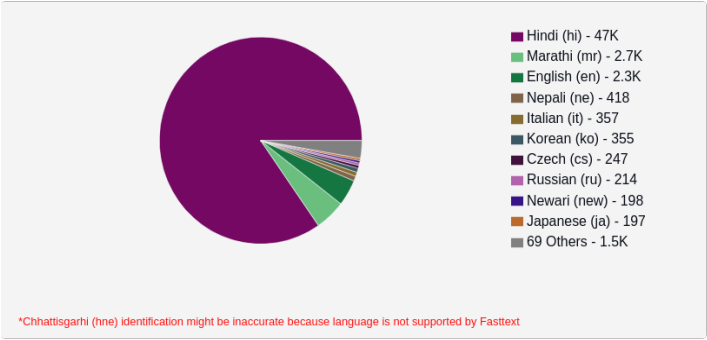


Documents by collection

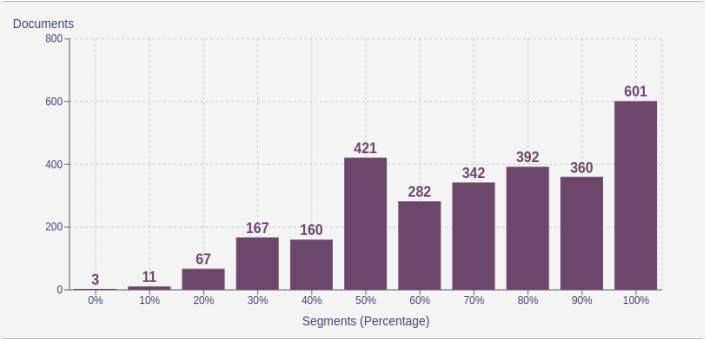


Language Distribution

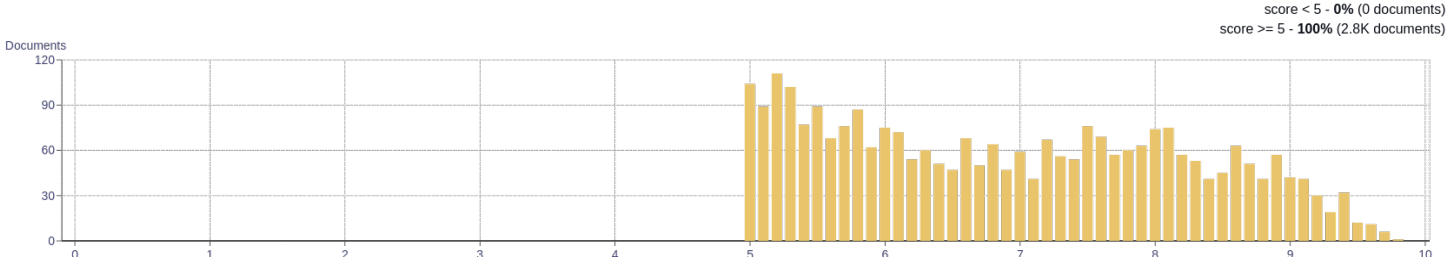
Number of segments



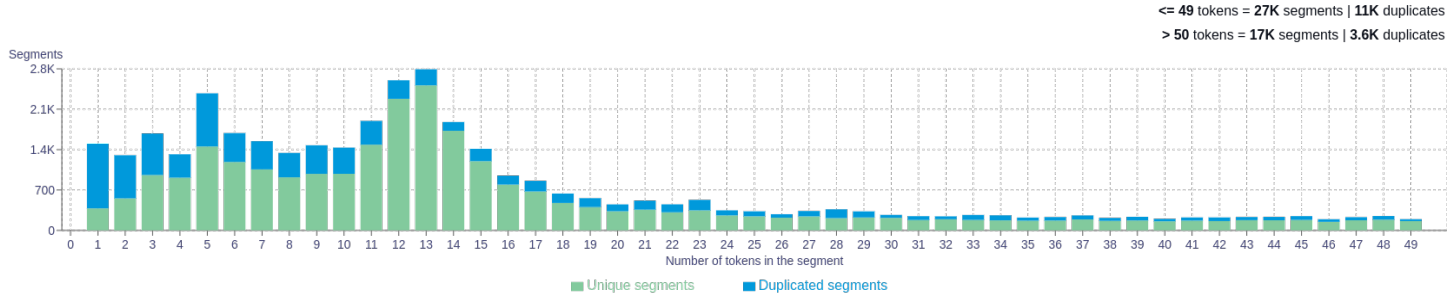
Percentage of segments in Chhattisgarhi (hne) inside documents



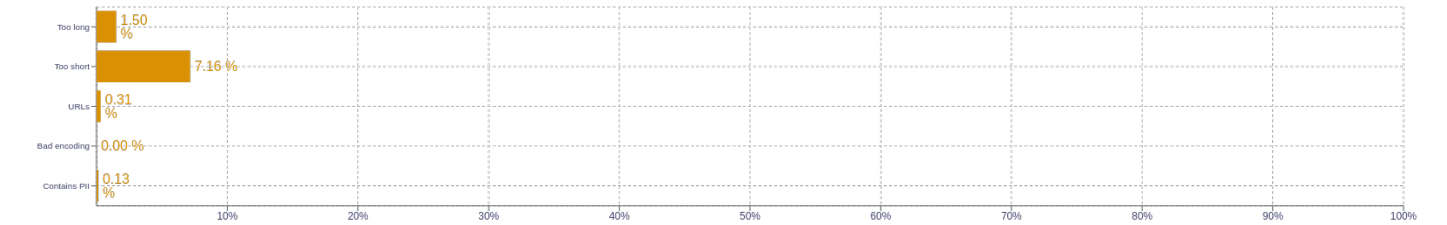
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	मन 33111ह 32170ले 30393हे 29654अउ 20944
2	देखे गए 6158बार देखे 6154ड्राइवर डाउनलोड 6154यो ह 3033मन ला 2046
3	बार देखे गए 6154keyword keyword keyword 1930name name name 1629insert column excel 714column excel shortcut 693
4	keyword keyword keyword keyword 1909name name name name 1624insert column excel shortcut 650gk quiz in hindi 309knowledge gk quiz in 306
5	keyword keyword keyword keyword keyword 1888name name name name name 1619knowledge gk quiz in hindi 306general knowledge gk quiz in 306 shortcut shortcut shortcut shortcut 255

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/mbanoni/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>