HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hin_Deva.jsonl.tsv	6/11/2025	Hindi (hi)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
13,651,945	267,232,818	149,937,458 (56.11 %)	9.6B	43,703,795,598	101.61 GB

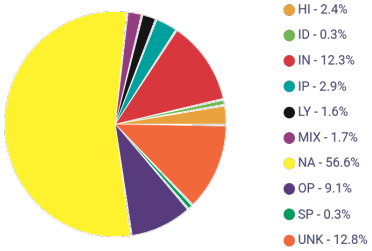
Top 10 domains

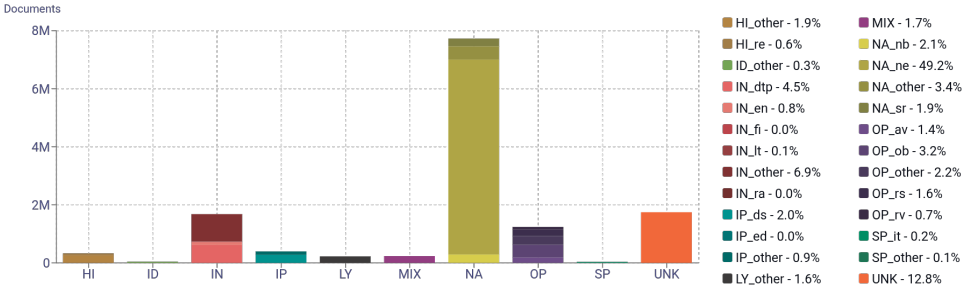
Domain	Docs	% of total
bhaskar.com	610K	4.47%
blogspot.com	435K	3.19%
blogspot.in	351K	2.57%
jagran.com	152K	1.11%
indiatimes.com	133K	0.97%
alibaba.com	128K	0.94%
amarujala.com	120K	0.88%
punjabkesari.in	110K	0.81%
wikipedia.org	95K	0.70%
webdunia.com	91K	0.67%

Top 10 TLDs

Domain	Docs	% of total
com	9.6M	70.08%
in	2.3M	16.61%
org	581K	4.26%
page	233K	1.71%
co.in	189K	1.38%
net	156K	1.14%
news	53K	0.39%
info	47K	0.34%
co	46K	0.34%
ae	39K	0.28%

Register labels

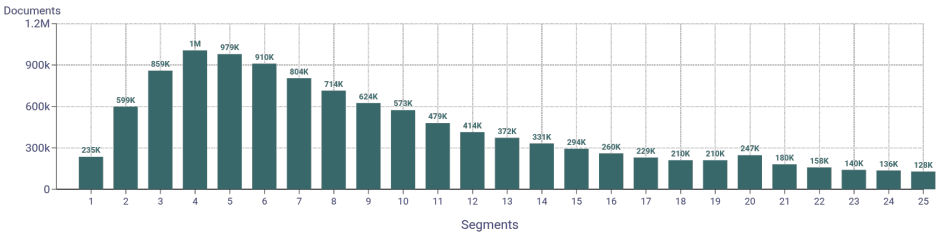




MT:6.7% | 913K Documents

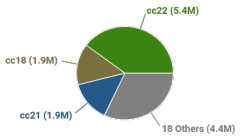
Documents size (in segments)

<= 25 segments **81.24%** (11M documents)
> 25 segments **18.76%** (2.6M documents)



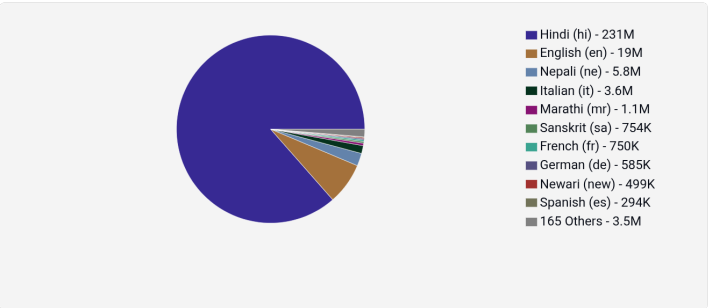
Documents by collection

CC = 74.77%
IA = 25.23%

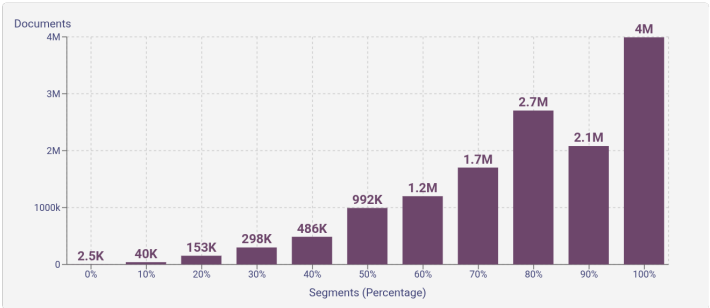


Language Distribution

Number of segments in the Hindi (hi) corpus

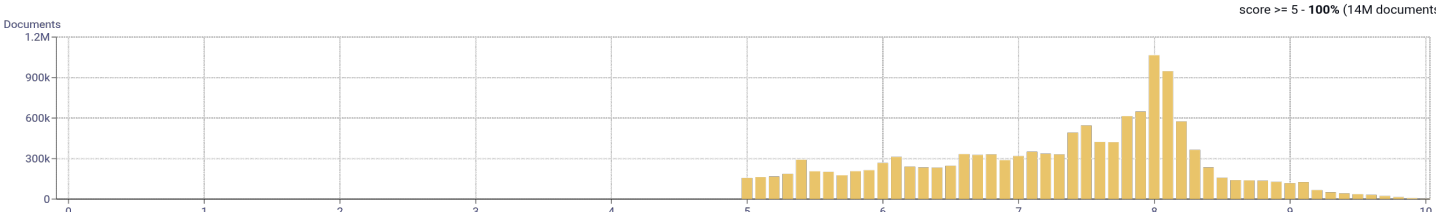


Percentage of segments in Hindi (hi) inside documents



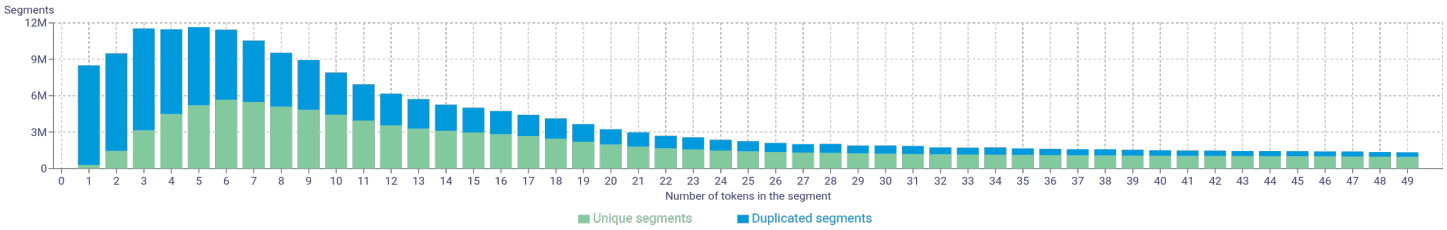
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (14M documents)

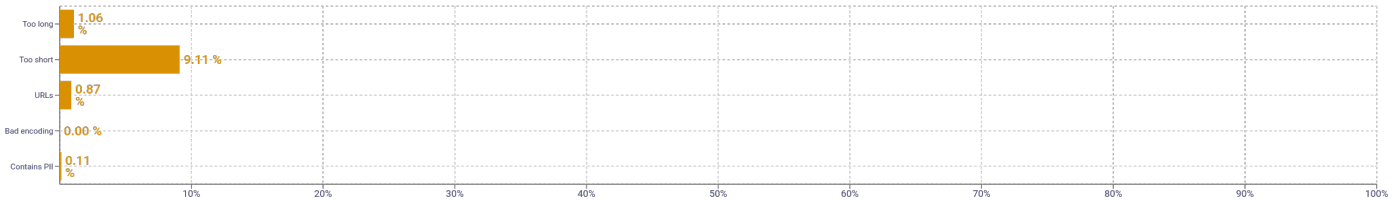


Segment length distribution by token

≤ 49 tokens = **104M** segments | **99M** duplicates
> 50 tokens = **65M** segments | **19M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	लेकिन | 15587369 · रही | 14673605 · जाता | 14341202 · अब | 14173322 · क्या | 13881956
2	in hindi | 2412309 · दी गई | 1677426 · उत्तर प्रदेश | 1536738 · नई दिल्ली | 1508348 · किए गए | 1481079
3	कम से कम | 989023 · मौत हो गई | 546181 · प्रधानमंत्री नरेंद्र मोदी | 519432 · योजना के तहत | 410823 · बारे में जानकारी | 339765
4	नाम से जाना जाता | 132924 · रूप में जाना जाता | 124506 · this website follows the | 115851 · website follows the dnpa | 115834 · the dnpa code of | 115830
5	this website follows the dnpa | 115833 · the dnpa code of ethics | 115830 · follows the dnpa code of | 115830 · website follows the dnpa code | 115829 · लोगों की मौत हो गई | 109655

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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av