HPLT Analytics report

HPLT Analytics

General overview

Corpus	Date	Language
hplt-v3-kas_Deva	9/18/2025	Kashmiri (ks)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
66	571	536 (93.87 %)	58K	251,421	594.53 KB

Top 10 domains

Domain	Docs	% of total
हिन्दीसेक्स.com	9	13.64%
xxxhindivideo.mobi	8	12.12%
wikipedia.org	5	7.58%
altanurag.mn	4	6.06%
xvidios.vip	3	4.55%
wikia.com	3	4.55%
thewiki.in	3	4.55%
nextnewmovies.com	3	4.55%
xnxn.mobi	2	3.03%
therightnews.in	2	3.03%

Top 10 TLDs

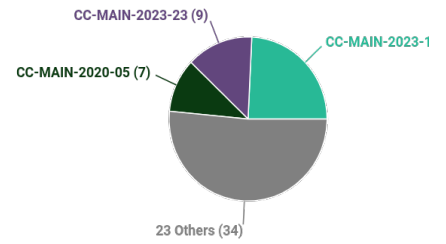| Domain	Docs	% of total
com	25	37.88%
mobi	12	18.18%
in	8	12.12%
org	7	10.61%
vip	4	6.06%
mn	4	6.06%
sn	2	3.03%
net	2	3.03%
ru	1	1.52%
click	1	1.52%

Documents size (in segments) ⓘ
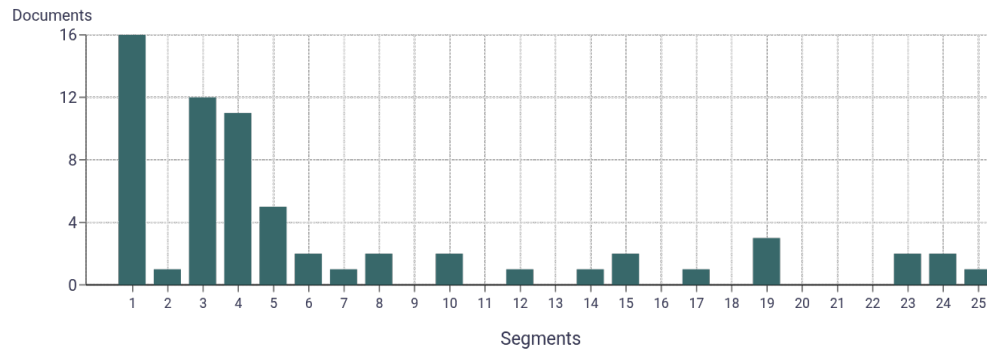
≤ 25 segments **98.48%** (65 documents)
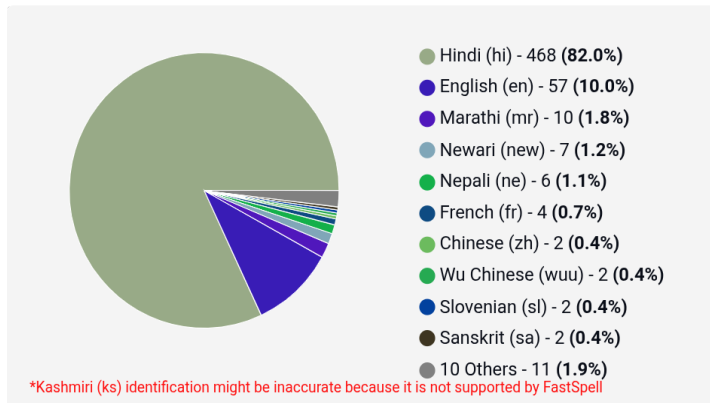> 25 segments **1.52%** (1 documents)

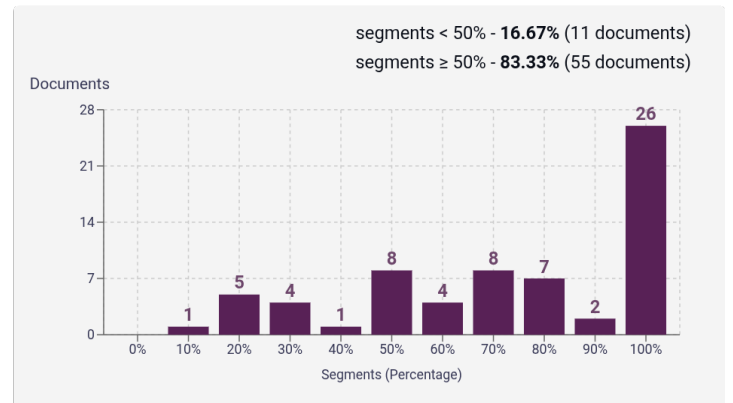## Document collections

CC = 95.45%
IA = 4.55%

CC-MAIN-2023-23 (9)
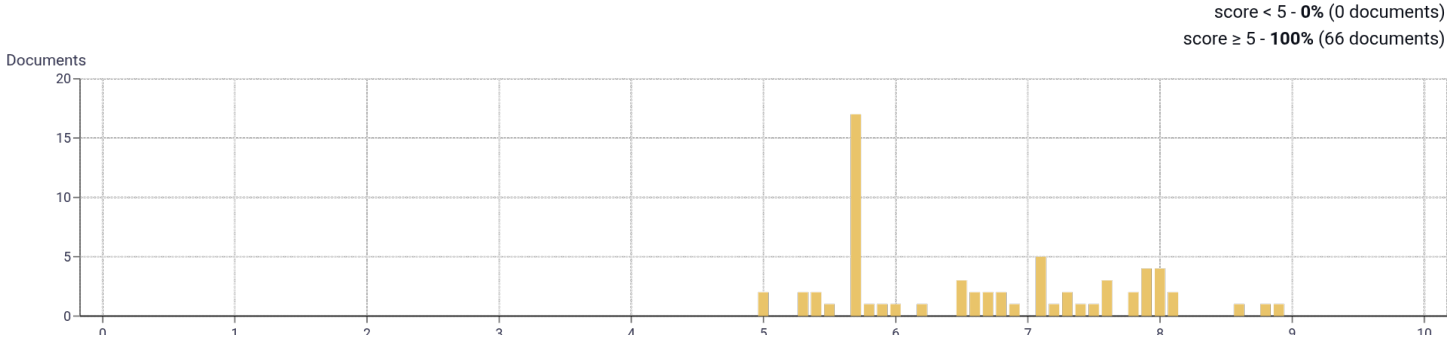CC-MAIN-2020-05 (7)
CC-MAIN-2023-1
23 Others (34)

Language Distribution

Number of segments in the Kashmiri (ks) corpus

- Hindi (hi) - 468 **(82.0%)**
- English (en) - 57 **(10.0%)**
- Marathi (mr) - 10 **(1.8%)**
- Newari (new) - 7 **(1.2%)**
- Nepali (ne) - 6 **(1.1%)**
- French (fr) - 4 **(0.7%)**
- Chinese (zh) - 2 **(0.4%)**
- Wu Chinese (wuu) - 2 **(0.4%)**
- Slovenian (sl) - 2 **(0.4%)**
- Sanskrit (sa) - 2 **(0.4%)**
- 10 Others - 11 **(1.9%)**

*Kashmiri (ks) identification might be inaccurate because it is not supported by FastSpell

Percentage of segments in Kashmiri (ks) inside documents

segments < 50% - **16.67%** (11 documents)
segments ≥ 50% - **83.33%** (55 documents)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (66 documents)

Documents

Segment length distribution by token

≤ 49 tokens = **327** segments | **34** duplicates
> 50 tokens = **244** segments | **3** duplicates

Segments

Segment noise distribution

Too long	**17.86%**
Too short	**8.93%**
URLs	**1.23%**
Bad encoding	**0.00%**
Contains PII	**0.18%**

■ % of corpus
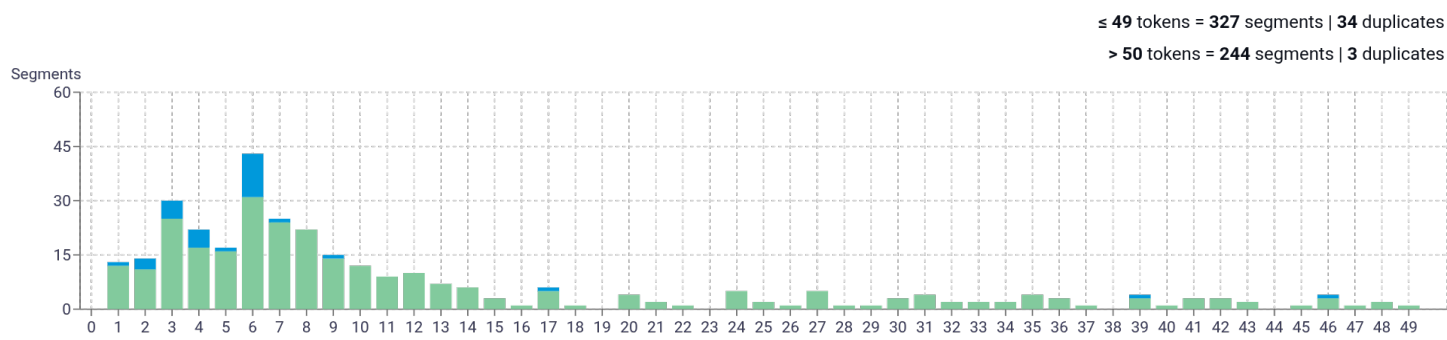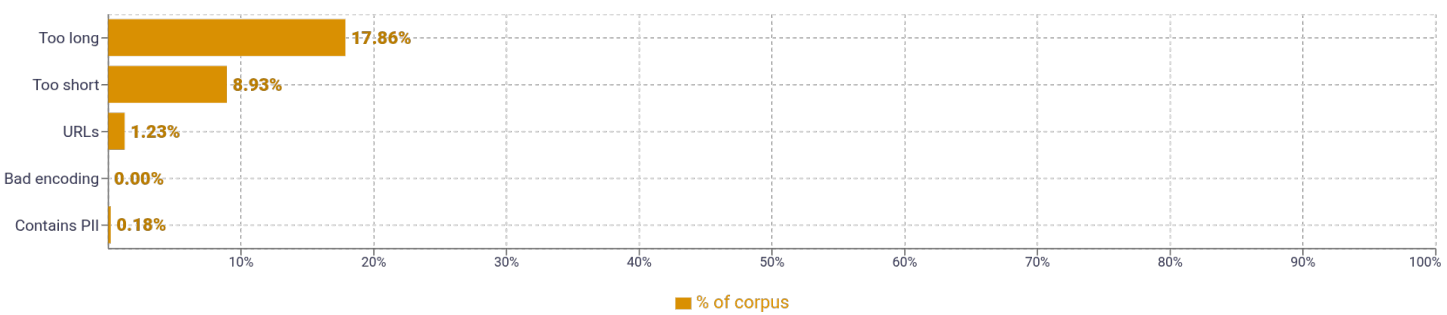
Frequent n-grams

SIZE	N-GRAMS				
1	क | 1,350	म | 816	सस | 803	ससससस | 740	ससस | 609
2	एक्स एक्स | 241	ससससस सस | 185	सस ससस | 138	एलश क | 134	ससससस ससससस | 126
3	एक्स एक्स एक्स | 147	ससससस ससससस सस | 33	ससससस सस ससस | 32	सस ससससससस सस | 32	सस ससस ससससस | 29
4	एक्स एक्स एक्स एक्स | 79	किस्स मिक्स किस्स मिक्स | 18	kiss mix kiss mix | 18	एक्स एक्स एक्स वीडियो | 14	सस ससससससस सस ससससस | 12
5	एक्स एक्स एक्स एक्स एक्स | 52	किस्स मिक्स किस्स मिक्स किस्स | 10	असल की नकल वाले गेम | 10	kiss mix kiss mix kiss | 10	वश्रा वश्रा किस्स मिक्स किस्स | 8

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.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	HI	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	**Informational persuasion**	IP	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	**Opinion**	**OP**
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	**Informational description**	IN	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				