

General overview

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

Volumes

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

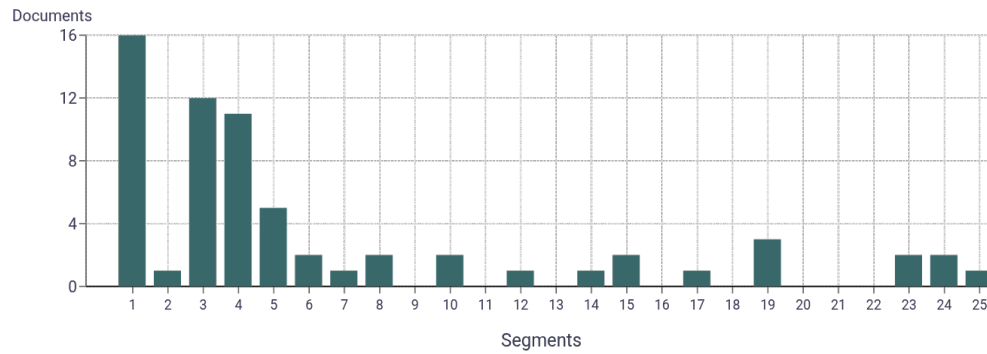
Top 10 domains

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%
xnxx.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%

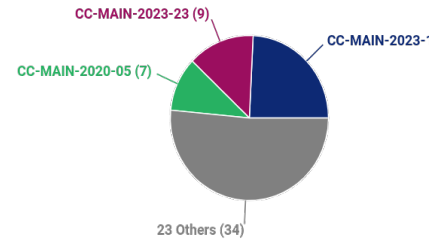
Documents size (in segments) ⓘ



≤ 25 segments **98.48%** (65 documents)
> 25 segments **1.52%** (1 documents)

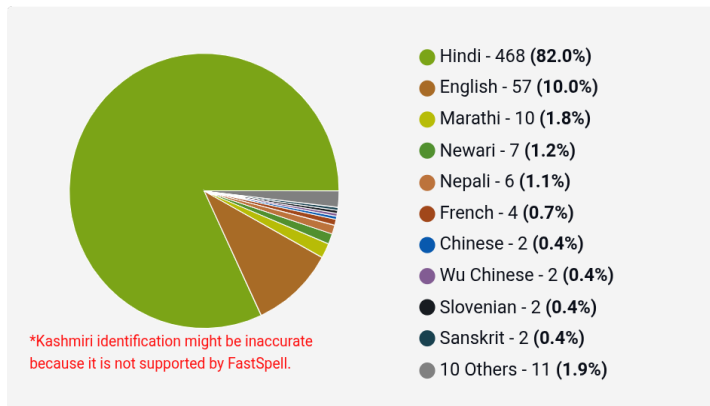
Document collections

CC = 95.45%
IA = 4.55%

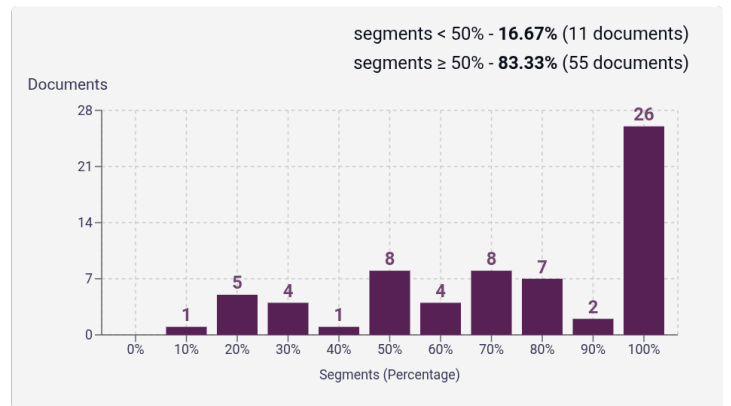


Language Distribution

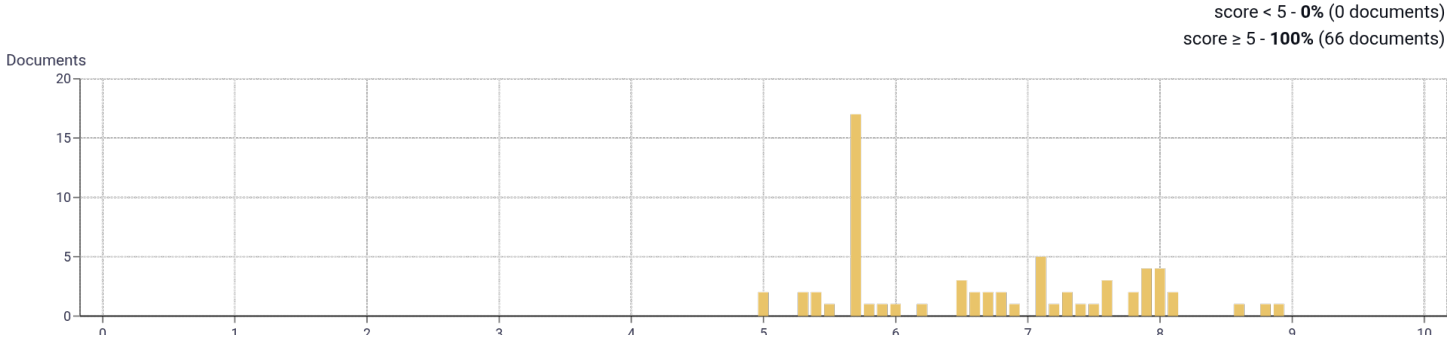
Number of segments in the Kashmiri corpus



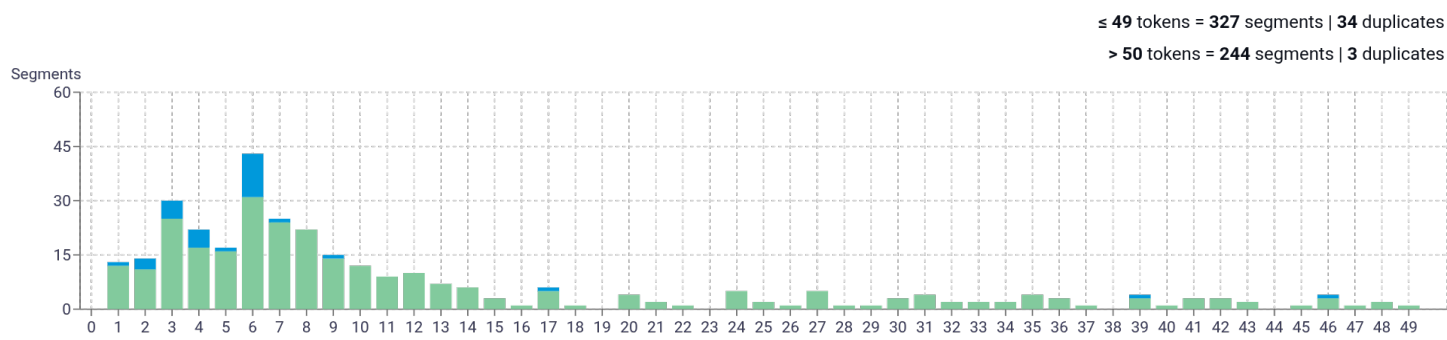
Percentage of segments in Kashmiri inside documents



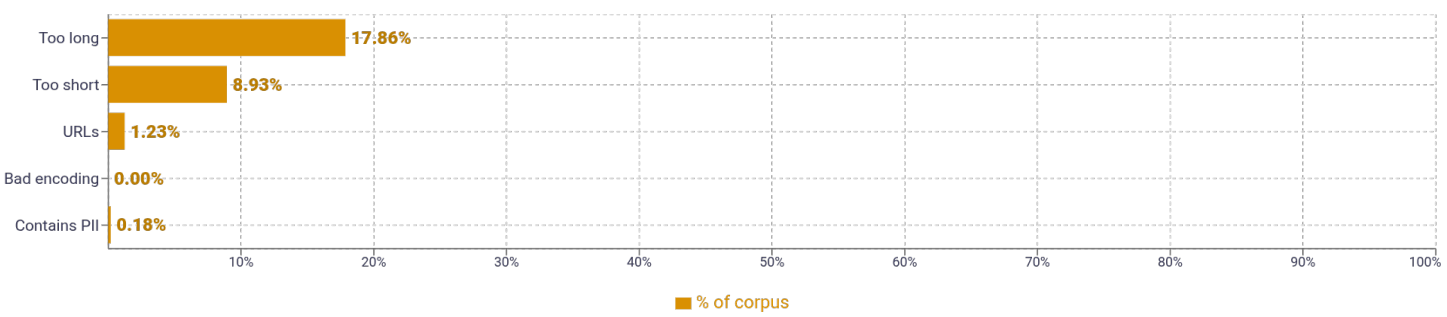
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	क 1,350 म 816 सस 803 सससस 740 ससस 609	
2	एक्स एक्स 241 सससस सस 185 सस ससस 138 एलश क 134 सससस सससस 126	
3	एक्स एक्स एक्स 147 सससस सससस सस 33 सससस सस ससस 32 सस सससससस सस 32 सस ससस सससस 29	
4	एक्स एक्स एक्स एक्स 79 किस्स मिक्स किस्स मिक्स 18 kiss mix kiss mix 18 एक्स एक्स एक्स वीडियो 14 सस सससससस सस सससस 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 वन्ना वन्ना किस्स मिक्स किस्स 8	

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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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>, , , 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	ntp
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				