

General overview

Corpus	Date	Language
hplt-v3-kas_Arab	9/17/2025	Kashmiri

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,067	25,433	20,583 (80.93 %)	707K	3,534,655	5.94 MB

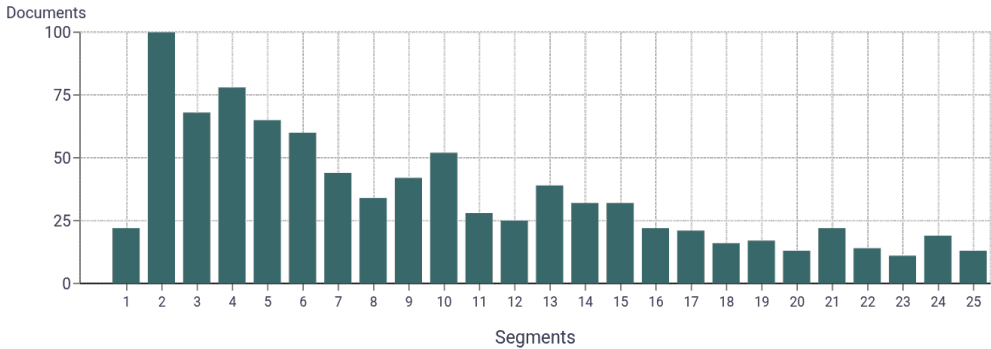
Top 10 domains

Domain	Docs	% of total
wikipedia.org	489	45.83%
muneeburrahman.com	147	13.78%
neabmagazine.com	106	9.93%
newschecker.in	78	7.31%
neabinternation...	43	4.03%
apsva.us	39	3.66%
koshurakhbar.com	34	3.19%
flightscanner.com	23	2.16%
wikimedia.org	20	1.87%
nidaekashmir.com	12	1.12%

Top 10 TLDs

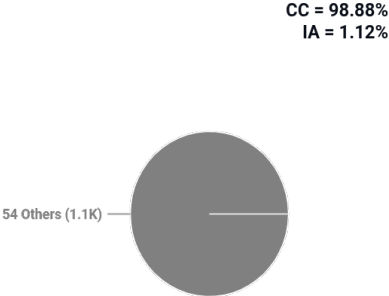
Domain	Docs	% of total
org	571	53.51%
com	348	32.61%
in	91	8.53%
us	39	3.66%
blog	5	0.47%
vn	2	0.19%
com.pk	2	0.19%
run	1	0.09%
ru	1	0.09%
org.in	1	0.09%

Documents size (in segments) ⓘ



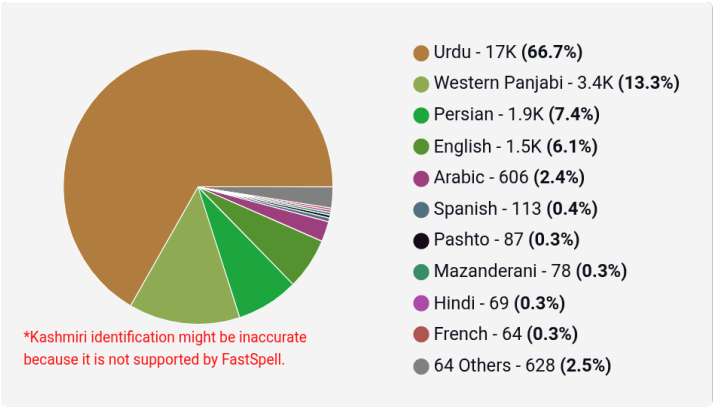
≤ 25 segments **83.32%** (889 documents)
> 25 segments **16.68%** (178 documents)

Document collections

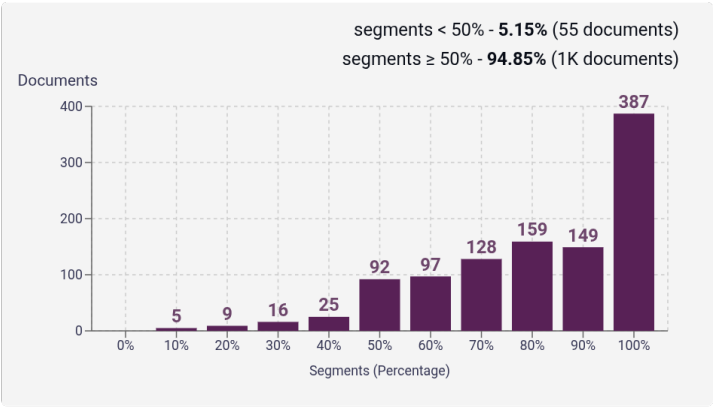


Language Distribution

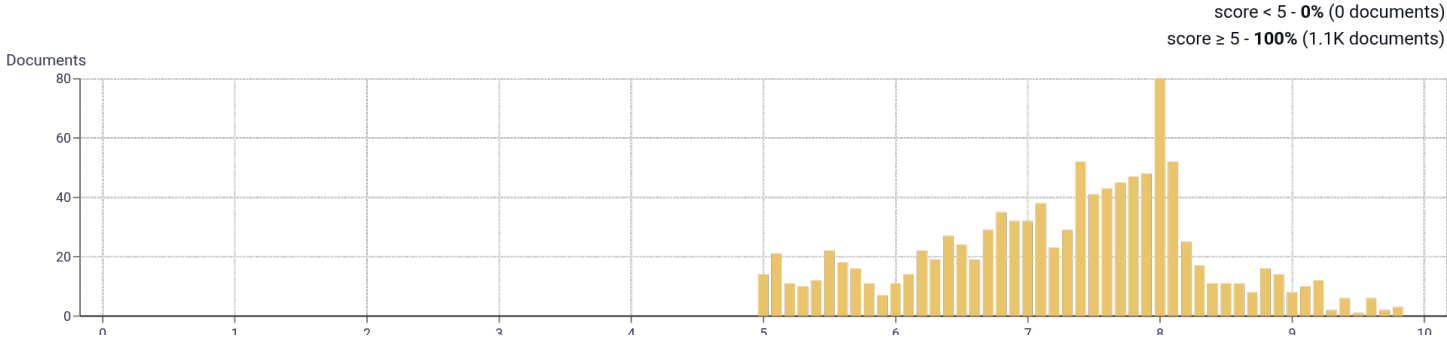
Number of segments in the Kashmiri corpus



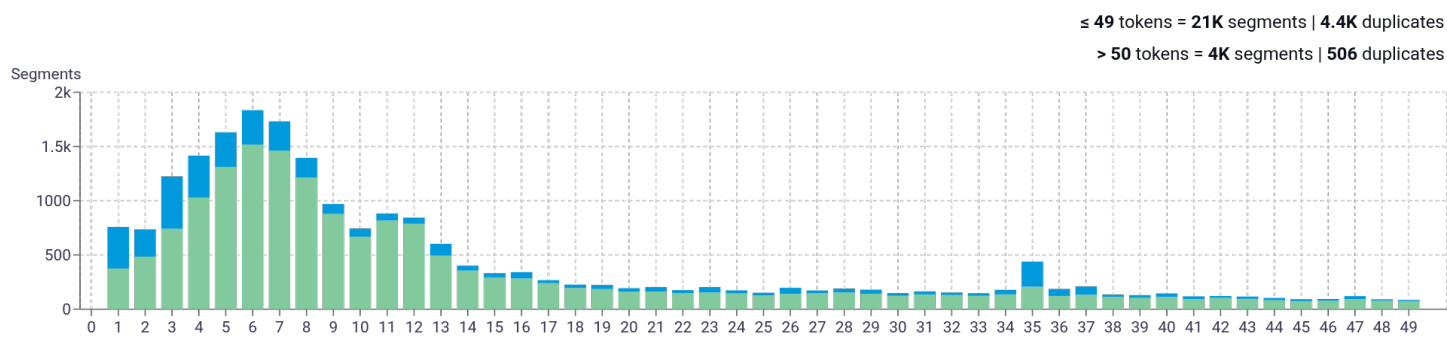
Percentage of segments in Kashmiri inside documents



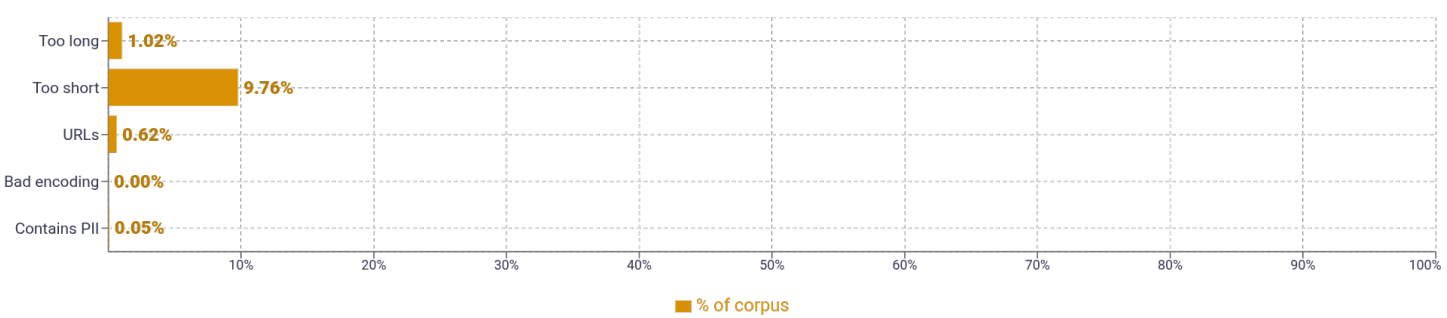
Distribution of documents by document score








Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	9,213 تر 8,079 منتر 7,205 چه 5,242 تر 4,951 فلفلر	
2	912 بی ایف 900 ایف ایل 889 چه تر 647 منتر چه 409 فن سیرو	
3	888 بی ایف ایل 399 بجاج فن سیرو 231 یو بی آئی 225 بجاج پء والیت 208 ای ایم آئی	
4	170 بی ایف ایل ک la la la la 160 بی ایف ایل پل 137 136 بی ایف ایل چه 116 بجاج فن سیرو پلیتفارم	
5	la la la la la 157 66 بی ایف ایل ک طرف archived from the original on 55 46 بجاج فن سیرو پلیت فارم 44 واجب الادا ای ایم آئی	

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				