

General overview

Corpus	Date	Language
hplt-v3-san_Deva	9/18/2025	Sanskrit

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
59,818	3,905,126	3,251,634 (83.27 %)	65M	425,468,015	1.07 GB

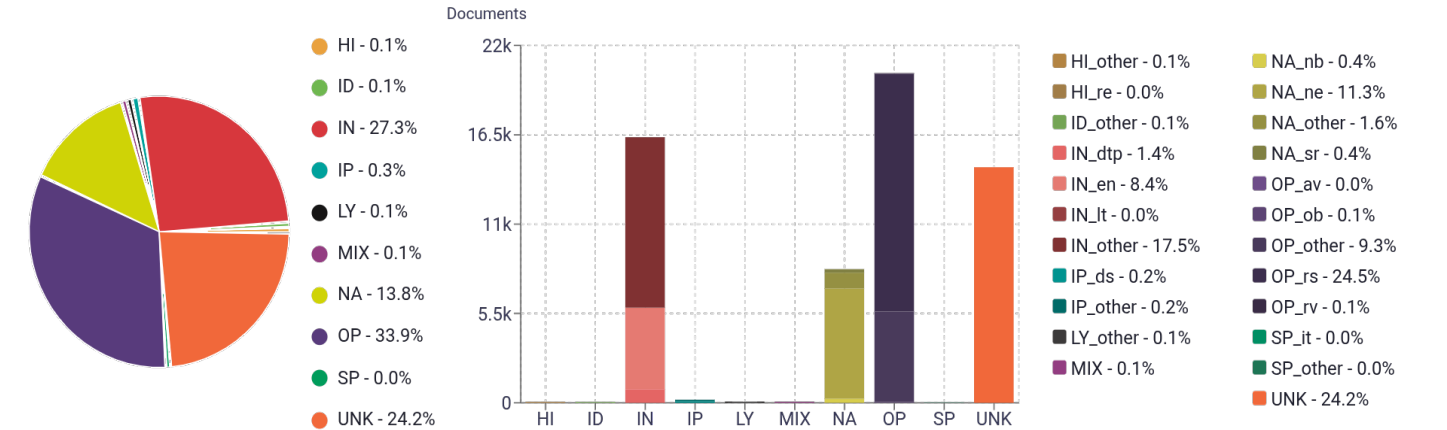
Top 10 domains

Domain	Docs	% of total
wikisource.org	13K	22.00%
wikipedia.org	5.1K	8.57%
sanskritdocumen...	3.7K	6.14%
blogspot.com	3.6K	6.10%
ashtadhyayi.com	2.6K	4.33%
sanskritvarta.in	2.2K	3.61%
avg-sanskrit.org	2K	3.35%
vedicscriptures.in	2K	3.30%
translital.org	1.7K	2.84%
indology.info	1.4K	2.38%

Top 10 TLDs

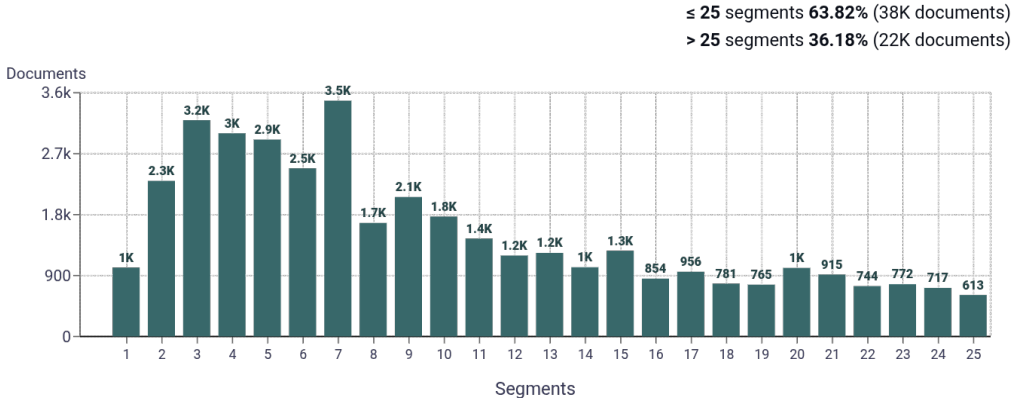
Domain	Docs	% of total
org	31K	51.05%
com	17K	28.79%
in	6.6K	11.02%
info	1.6K	2.73%
gov.in	1.3K	2.11%
co.in	751	1.26%
net	375	0.63%
ac.in	323	0.54%
blog	167	0.28%
de	160	0.27%

Register labels

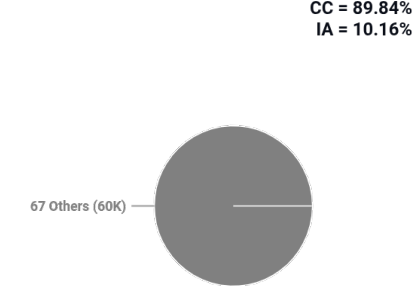


MT:0.6% | 373 Documents

Documents size (in segments) ⓘ

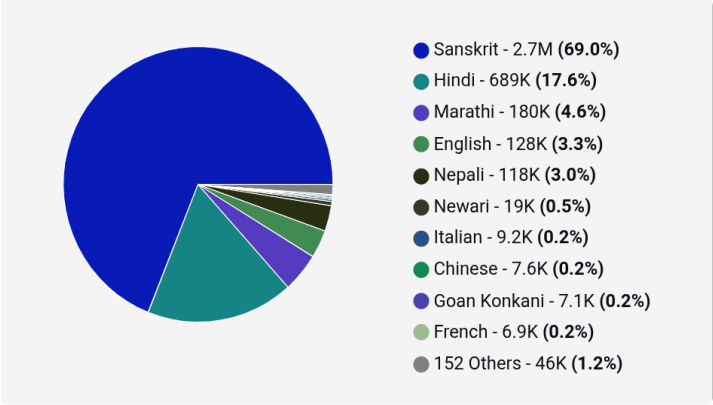


Document collections

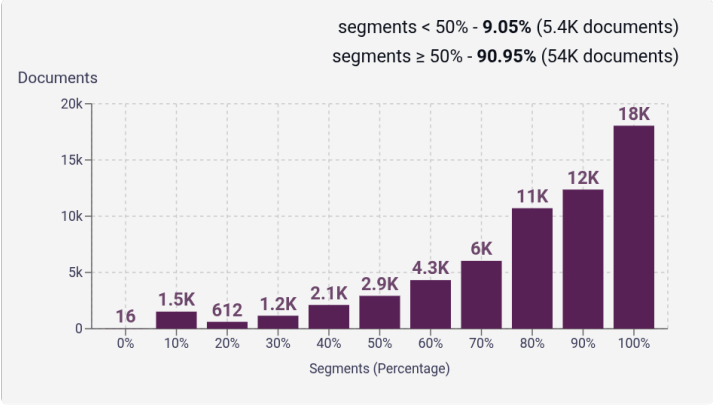


Language Distribution

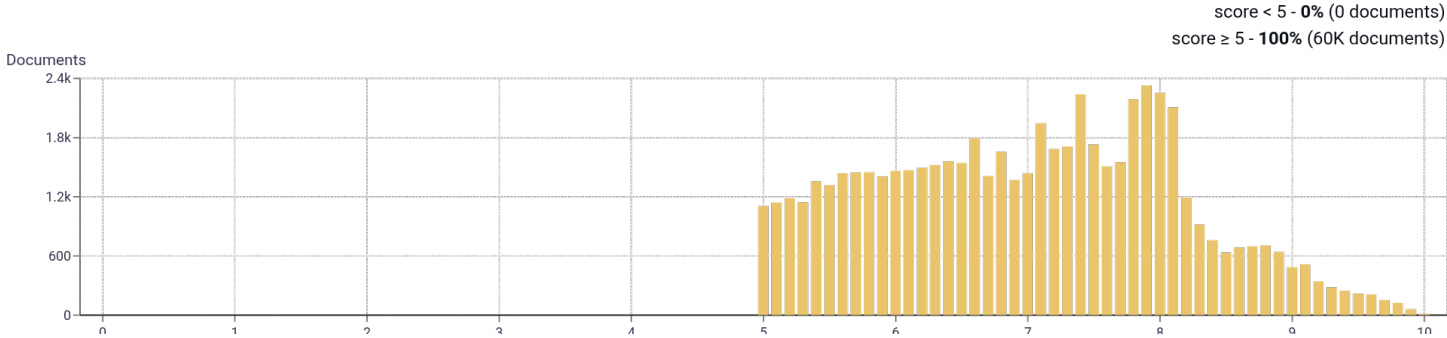
Number of segments in the Sanskrit corpus



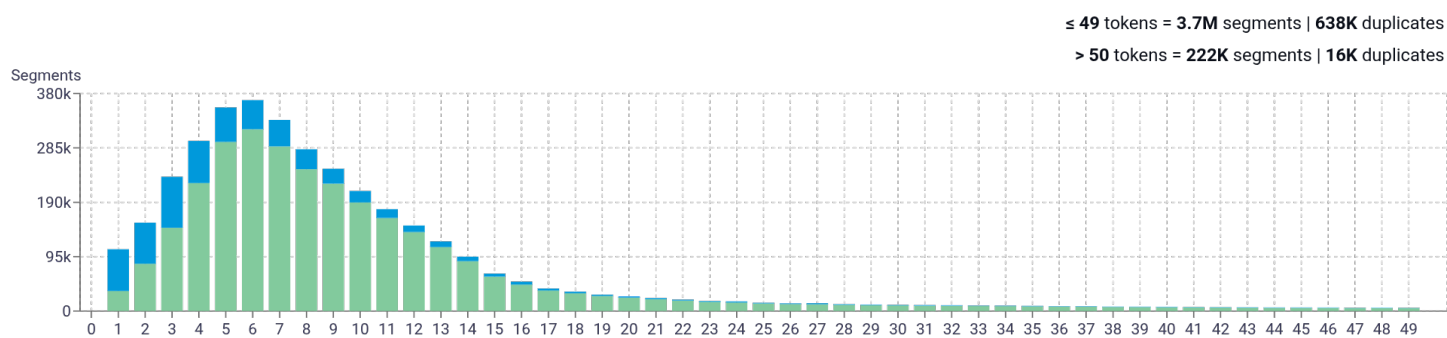
Percentage of segments in Sanskrit inside documents



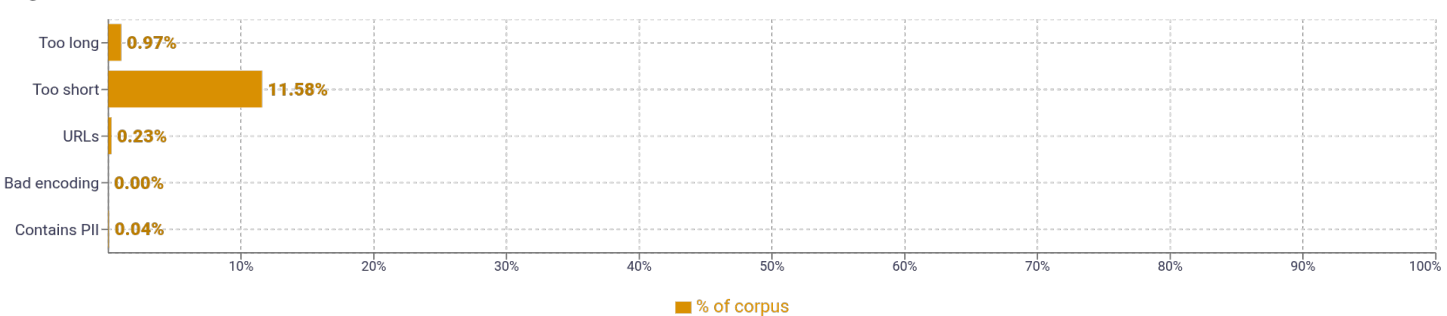
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	इति 587,474न 546,805स 187,416तथा 132,010एवं 91,302	
2	तमे वर्षे 8,604of the 7,411इति भावः 6,968in sanskrit 5,315proofing by 5,216	
3	rig veda book 4,418य एवं वेद 3,492mahabharata in sanskrit 2,022jump to navigation 2,017jump to search 2,016	
4	the mahabharata in sanskrit 2,015mahabharata in sanskrit book 1,995will be updated soon 1,848use feedback link below 1,401to provide the text 1,401	
5	the mahabharata in sanskrit book 1,995use feedback link below to 1,401to provide the text if 1,401the text if you have 1,401provide the text if you 1,401	

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	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				