HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
san_Deva.jsonl.tsv	9/6/2024	Sanskrit (sa)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
54,911	3,281,167	1,723,439 (52.53 %)	55M	355,931,245	911.79 MB

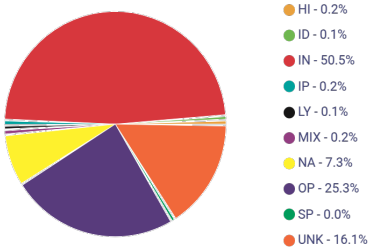
Top 10 domains

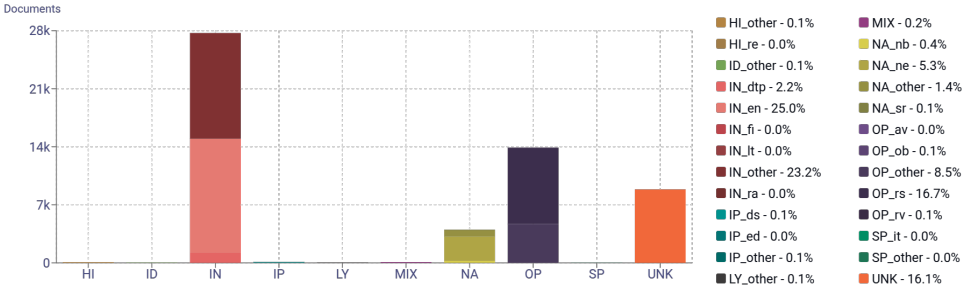
Domain	Docs	% of total
wikipedia.org	16K	29.27%
wikisource.org	7.2K	13.17%
sanskritdocumen...	5.8K	10.62%
blogspot.com	3.7K	6.65%
ashtadhyayi.com	3.2K	5.81%
avg-sanskrit.org	2.3K	4.15%
indology.info	1.7K	3.16%
transliteral.org	1.6K	2.96%
wikiquote.org	922	1.68%
upasanayoga.org	803	1.46%

Top 10 TLDs

Domain	Docs	% of total
org	38K	68.46%
com	12K	21.49%
info	1.8K	3.30%
in	1.4K	2.48%
co.in	555	1.01%
net	421	0.77%
gov.in	305	0.56%
ac.in	201	0.37%
page	176	0.32%
blog	151	0.27%

Register labels

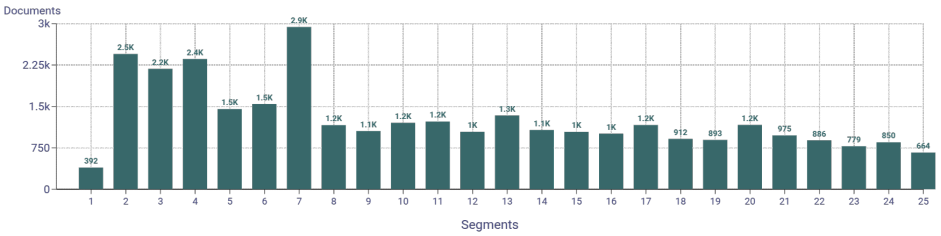




MT:0.2% | 122 Documents

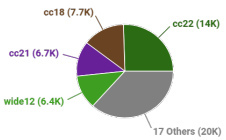
Documents size (in segments)

<= 25 segments **57.81%** (32K documents)
> 25 segments **42.19%** (23K documents)



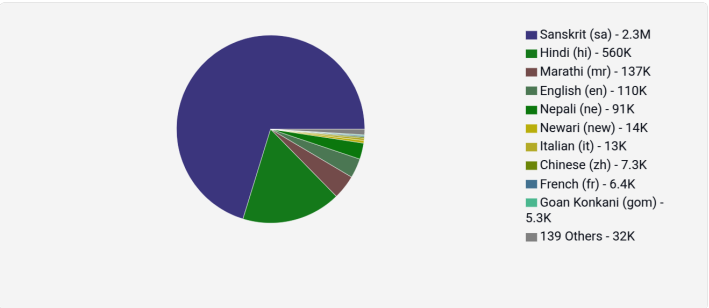
Documents by collection

CC = 64.95%
IA = 35.05%

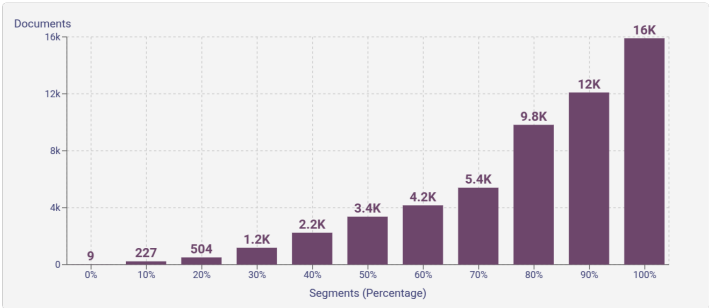


Language Distribution

Number of segments in the Sanskrit (sa) corpus

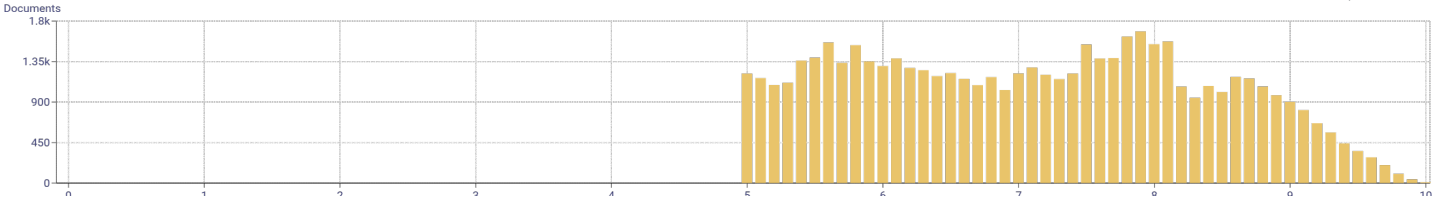


Percentage of segments in Sanskrit (sa) inside documents



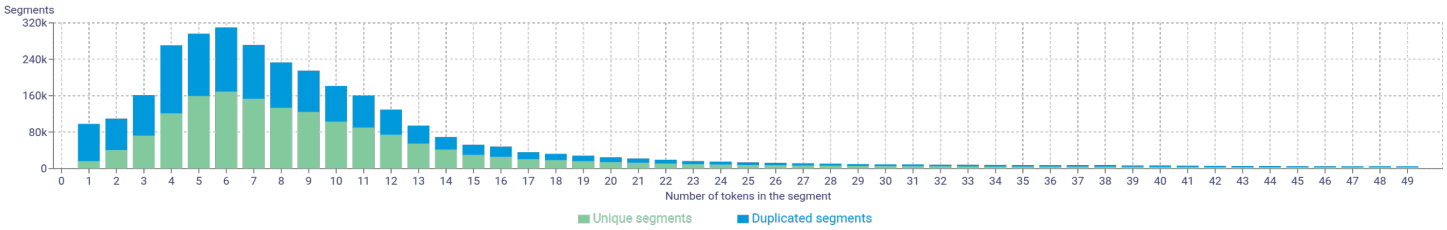
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (55K documents)

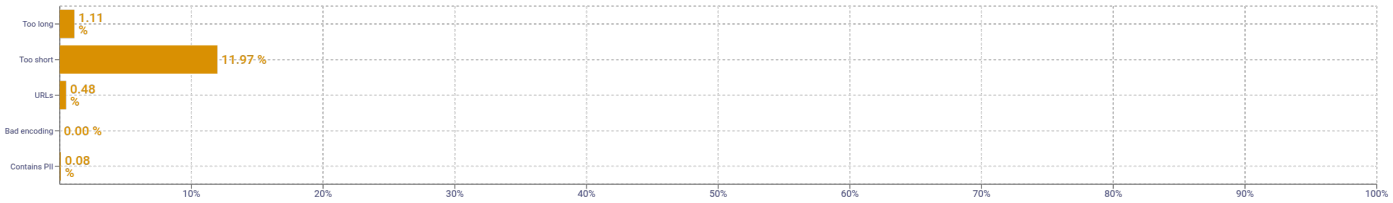


Segment length distribution by token

≤ 49 tokens = **1.6M** segments | **1.5M** duplicates
> 50 tokens = **202K** segments | **96K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	इति | 517838 · न | 390950 · स | 147621 · तथा | 109088 · सम्पादयतु | 81267
2	तमे वर्षे | 22363 · of the | 7789 · नमो नमः | 4588 · in the | 4294 · य एवं | 3912
3	और प्रशिक्षण परिषद् | 3778 · शैक्षिक अनुसंधान और | 3724 · राष्ट्रीय शैक्षिक अनुसंधान | 3724 · अनुसंधान और प्रशिक्षण | 3724 · य एवं वेद | 2728
4	शैक्षिक अनुसंधान और प्रशिक्षण | 3724 · राष्ट्रीय शैक्षिक अनुसंधान और | 3724 · अनुसंधान और प्रशिक्षण परिषद् | 3724 · सहस्रनामस्तोत्र यात आलेले नाम | 2214 · श्री विष्णु सहस्रनामस्तोत्र यात | 2214
5	शैक्षिक अनुसंधान और प्रशिक्षण परिषद् | 3724 · राष्ट्रीय शैक्षिक अनुसंधान और प्रशिक्षण | 3724 · श्री विष्णु सहस्रनामस्तोत्र यात आलेले | 2214 · विष्णु सहस्रनामस्तोत्र यात आलेले नाम | 2214 · use feedback link below to | 1700

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