HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-guj_Gujr	9/18/2025	Gujarati (gu)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
3,460,235	46,733,675	34,253,918 (73.30 %)	1.6B	8,343,071,750	19.92 GB

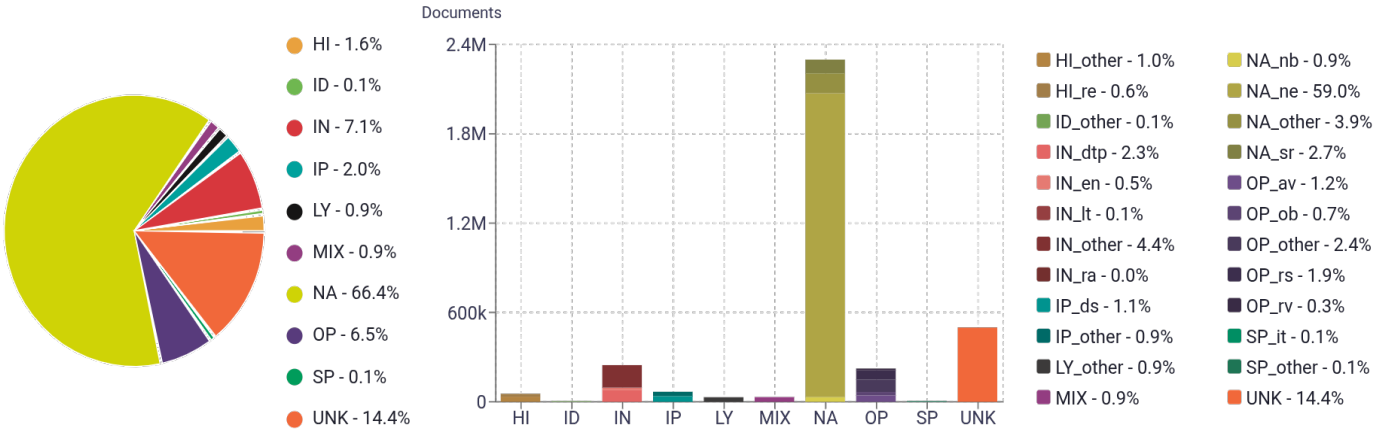
Top 10 domains

Domain	Docs	% of total
divyabhaskar.co.in	269K	7.76%
akilanews.com	198K	5.73%
news18.com	117K	3.39%
gujaratimidday.com	71K	2.06%
oneindia.com	68K	1.97%
tv9gujarati.com	60K	1.75%
iamgujarat.com	55K	1.59%
wordpress.com	51K	1.48%
chitralekha.com	51K	1.47%
gstv.in	47K	1.37%

Top 10 TLDs

Domain	Docs	% of total
com	2.3M	66.55%
in	560K	16.20%
co.in	324K	9.36%
org	89K	2.58%
net	64K	1.86%
news	18K	0.52%
online	9.1K	0.26%
club	6.9K	0.20%
co.uk	6.3K	0.18%
info	6.1K	0.18%

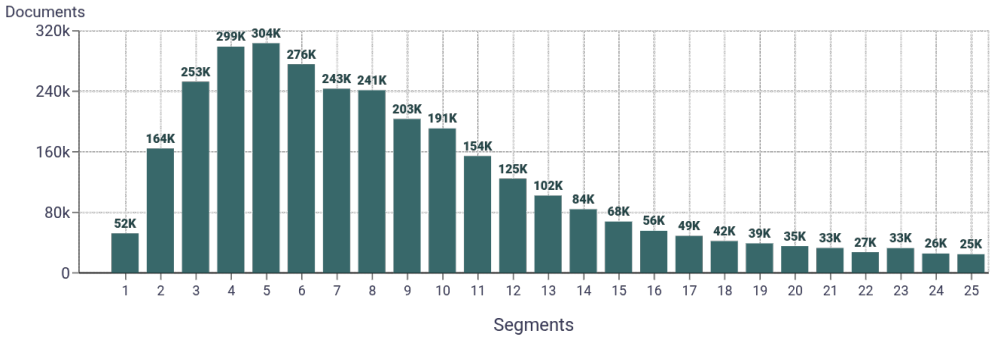
Register labels



MT:7.9% | 275K Documents

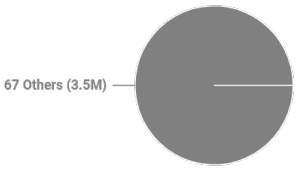
Documents size (in segments)

≤ 25 segments **90.31%** (3.1M documents)
> 25 segments **9.69%** (335K documents)



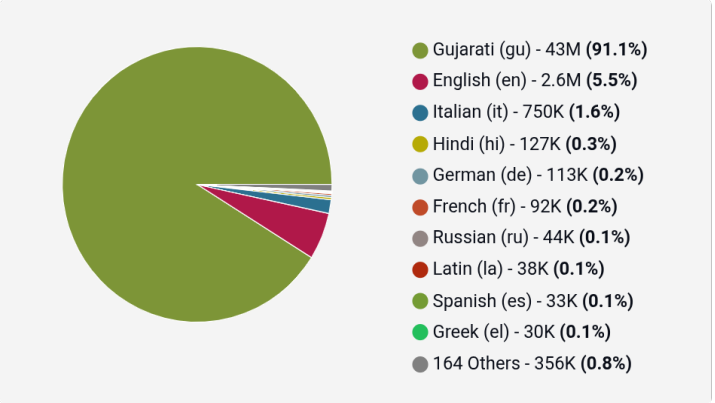
Document collections

CC = 95.92%
IA = 4.08%

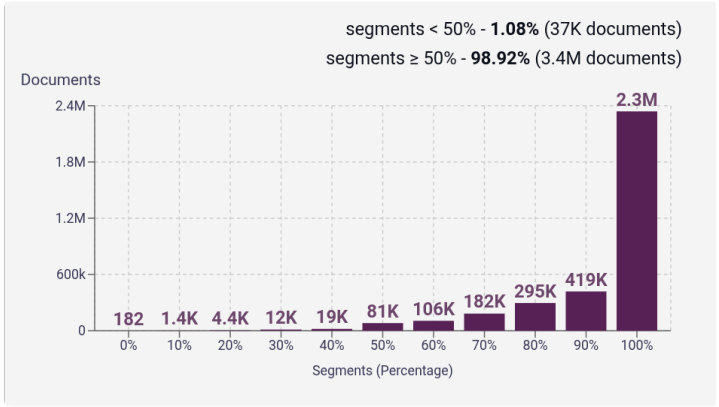


Language Distribution

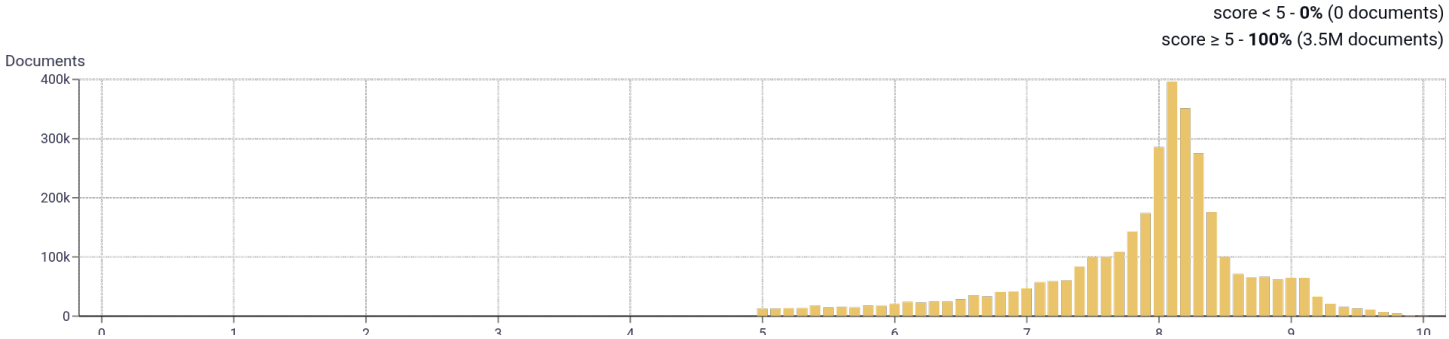
Number of segments in the Gujarati (gu) corpus



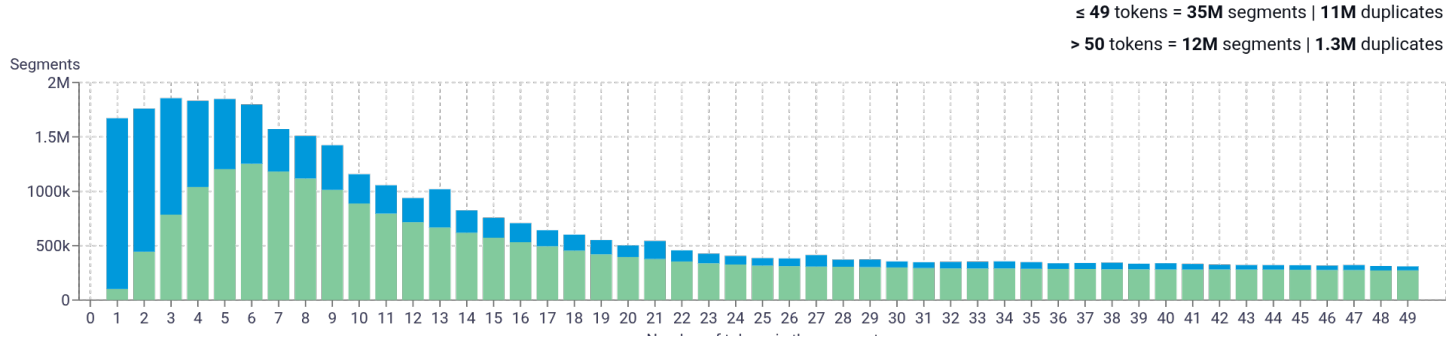
Percentage of segments in Gujarati (gu) inside documents



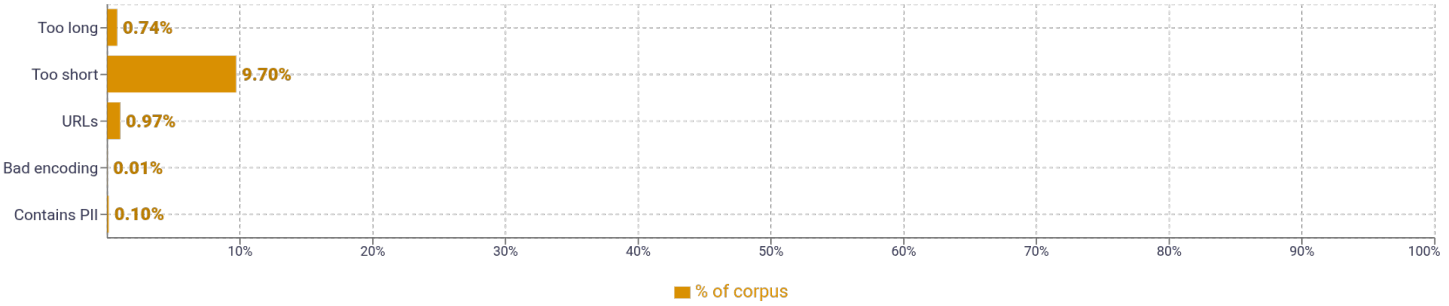
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	સાથે | 7,410,507 · હતો | 4,834,122 · કરવામાં | 4,095,437 · દ્વારા | 3,363,750 · તમારા | 2,414,108
2	આવ્યો હતો | 471,690 · કરવામાં આવ્યું | 399,076 · કરવામાં આવશે | 395,538 · જોવા મળી | 393,602 · કરવામાં આવ્યો | 387,345
3	કરવામાં આવ્યો હતો | 141,717 · તપાસ હાથ ધરી | 81,006 · આયોજન કરવામાં આવ્યું | 76,923 · all rights reserved | 69,616 · db corp ltd | 69,087
4	website follows the dnpa | 67,182 · this website follows the | 67,182 · the dnpa code of | 67,182 · follows the dnpa code | 67,182 · dnpa code of ethics | 67,182
5	website follows the dnpa code | 67,182 · this website follows the dnpa | 67,182 · the dnpa code of ethics | 67,182 · follows the dnpa code of | 67,182 · વગર સમાચાર વાંચવા ઇન્સ્ટોલ કરો | 54,337

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				