HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
guj_Gujr.jsonl.tsv	9/16/2024	Gujarati (gu)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,134,252	20,639,718	11,424,183 (55.35 %)	667M	3,366,421,654	7.99 GB

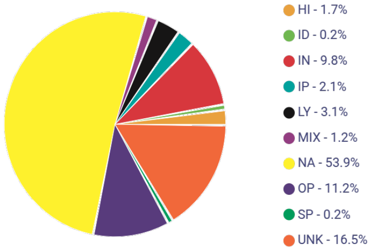
Top 10 domains

Domain	Docs	% of total
divyabhaskar.co.in	83K	7.35%
wordpress.com	48K	4.25%
news18.com	44K	3.92%
oneindia.com	29K	2.58%
sandesh.com	27K	2.42%
wikipedia.org	24K	2.16%
gujjurocks.in	23K	2.02%
chitralekha.com	18K	1.58%
webdunia.com	17K	1.48%
vtvgujarati.com	16K	1.38%

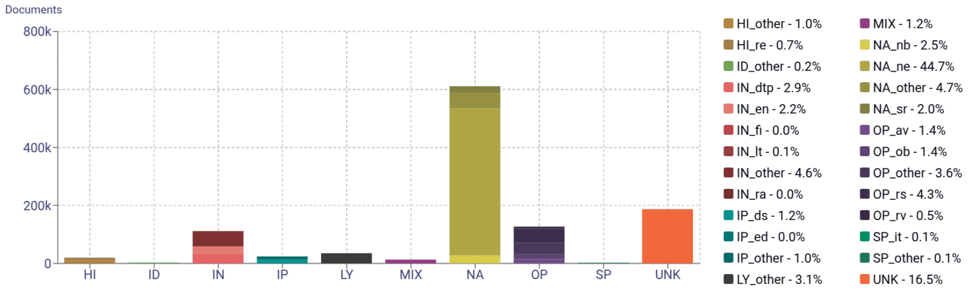
Top 10 TLDs

Domain	Docs	% of total
com	727K	64.10%
in	172K	15.18%
co.in	101K	8.92%
org	76K	6.69%
net	19K	1.67%
news	4.7K	0.41%
app	3K	0.26%
online	2.8K	0.24%
gov.in	2.4K	0.21%
live	1.4K	0.12%

Register labels

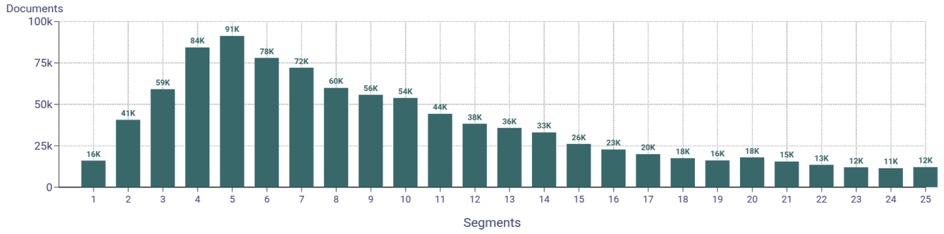


🤖 **MT**:8.4% | 96K Documents



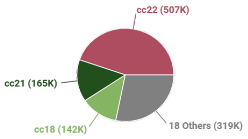
Documents size (in segments)

<= 25 segments **83.56%** (948K documents)
> 25 segments **16.44%** (186K documents)



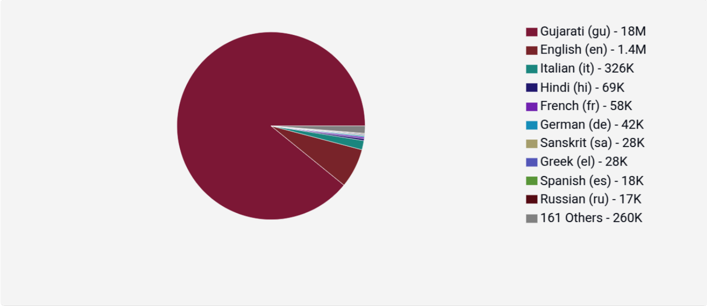
Documents by collection

CC = 77.77%
IA = 22.23%

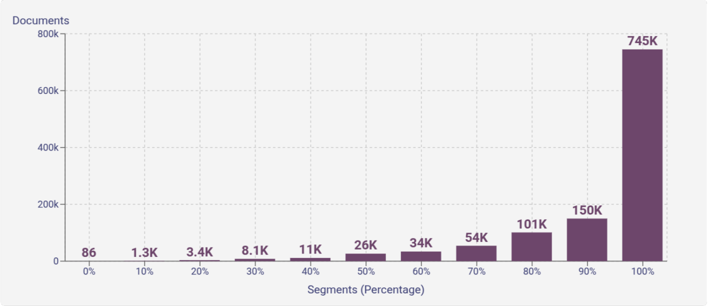


Language Distribution

Number of segments in the Gujarati (gu) corpus

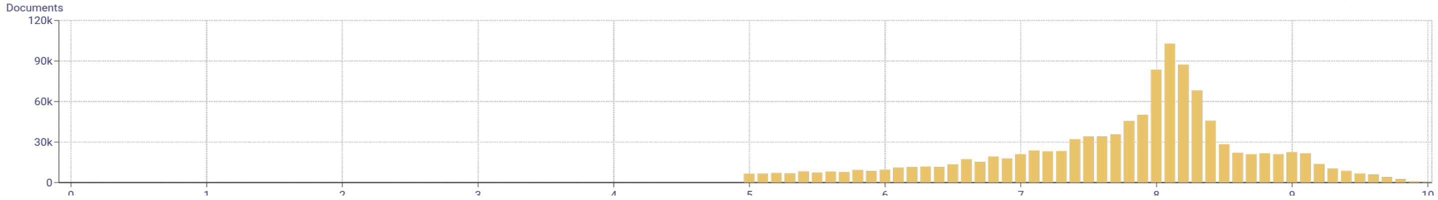


Percentage of segments in Gujarati (gu) inside documents



Distribution of documents by document score

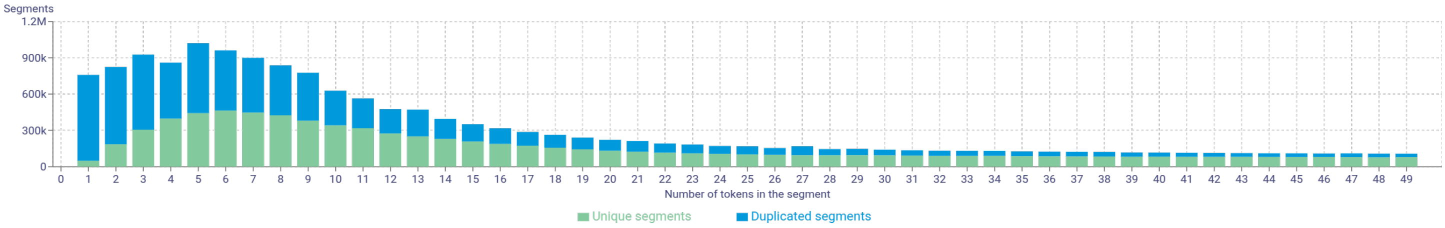
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (1.1M documents)



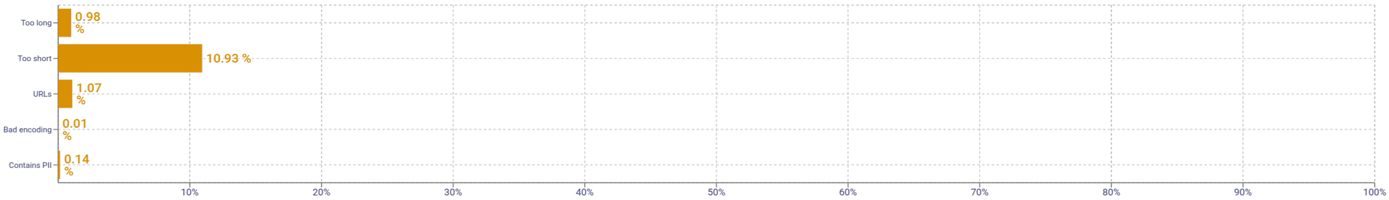
Segment length distribution by token

≤ **49** tokens = **8.1M** segments | **7.9M** duplicates

> **50** tokens = **4.6M** segments | **1.3M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	સાથે | 2721183 હતો | 1796456 કરવામાં | 1275130 કરવા | 1137108 દ્વારા | 1131630
2	ફેરફાર કરો | 177866 આવ્યો હતો | 143408 જોવા મળે | 130076 કરવામાં આવ્યું | 122479 કરવામાં આવ્યો | 119799
3	all rights reserved | 46726 db corp ltd | 46044 code of ethics | 46018 website follows the | 46017 this website follows | 46017
4	website follows the dnpa | 46017 this website follows the | 46017 the dnpa code of | 46017 follows the dnpa code | 46017 dnpa code of ethics | 46017
5	website follows the dnpa code | 46017 this website follows the dnpa | 46017 the dnpa code of ethics | 46017 follows the dnpa code of | 46017 સહિત વધુ સમાચાર વાંચો ન્યૂઝ18 | 20648

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