HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
gu_1.jsonl.tsv	3/17/2024	Gujarati (gu)

Volumes

Docs	Segments	Unique segments	Tokens	Size
264,816	32,742,913	31,437 (0.10 %)	369M	3.79 GB

Type-Token Ratio

Gujarati (gu)
0.01

Top 10 domains

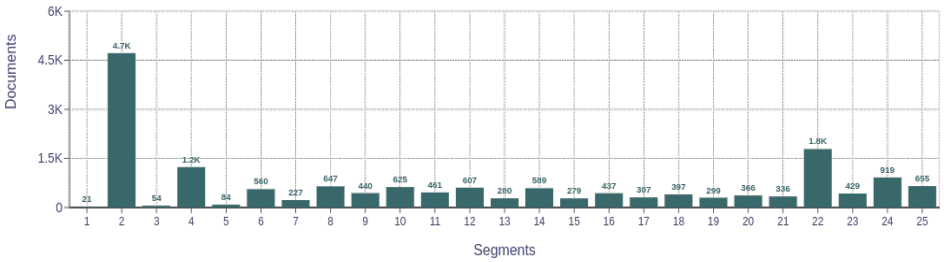
Domain	Docs	% of total
divyabhaskar.co.in	12K	4.44
iamgujarat.com	11K	4.02
gujjurocks.in	10K	3.87
news18.com	10K	3.83
blogspot.in	9.9K	3.74
akilanews.com	9.1K	3.44
indiatimes.com	6.7K	2.52
wordpress.com	5.7K	2.16
oneindia.com	3.9K	1.49
aksharnaad.com	3.8K	1.44

Top 10 TLDs

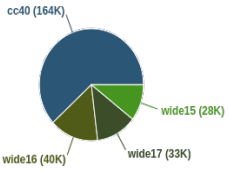
Domain	Docs	% of total
com	169K	63.73
in	51K	19.21
co.in	18K	6.71
org	15K	5.56
net	3K	1.13
app	953	0.36
news	941	0.36
gov.in	710	0.27
info	614	0.23
com.au	523	0.20

Documents size (in segments)

<= 25 segments **6.33%** (17K documents)
> 25 segments **93.67%** (248K documents)

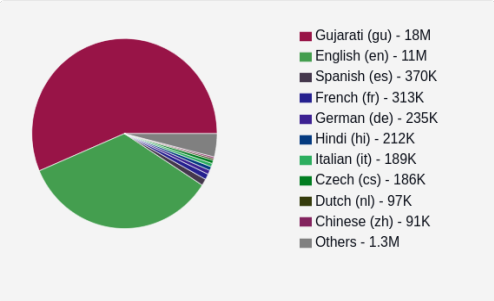


Documents by collection

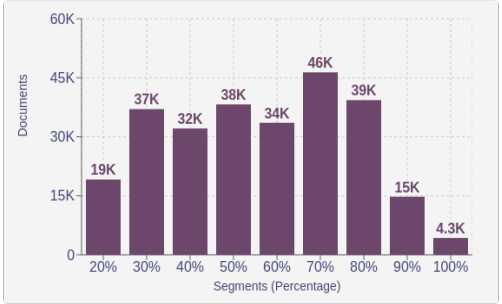


Language Distribution

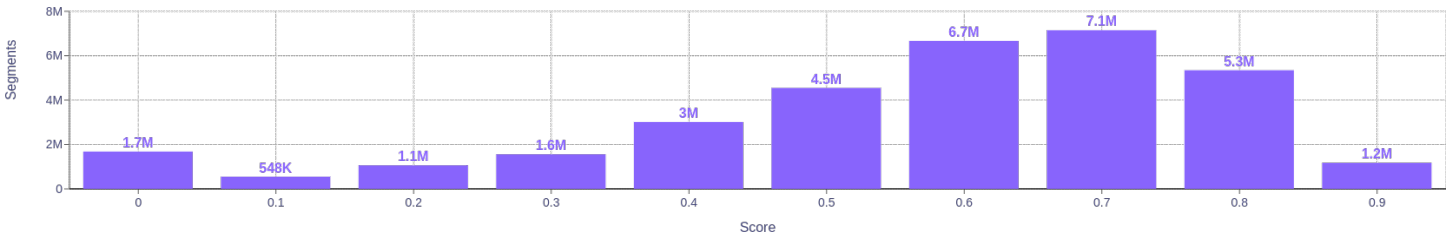
Number of segments



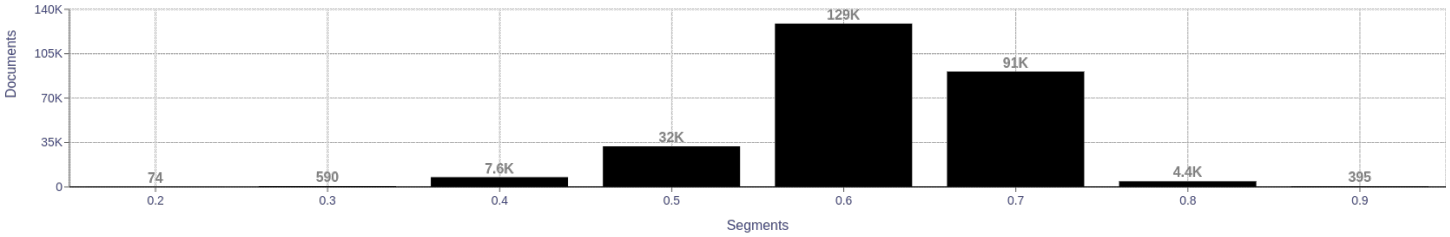
Percentage of segments in Gujarati (gu) inside documents



Distribution of segments by fluency score

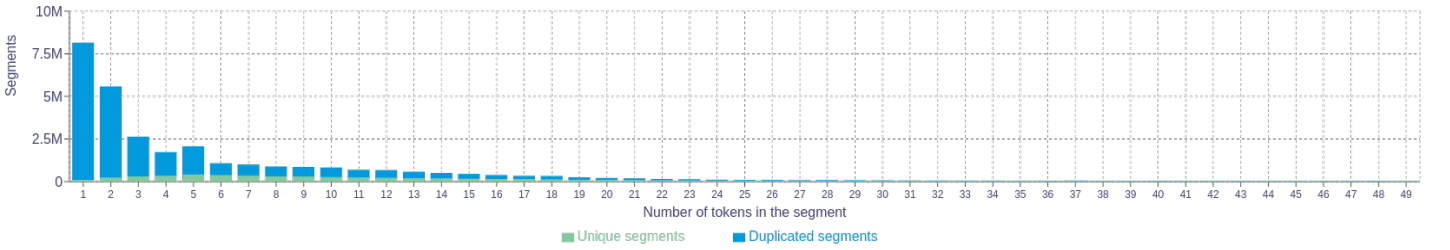


Distribution of documents by average fluency score

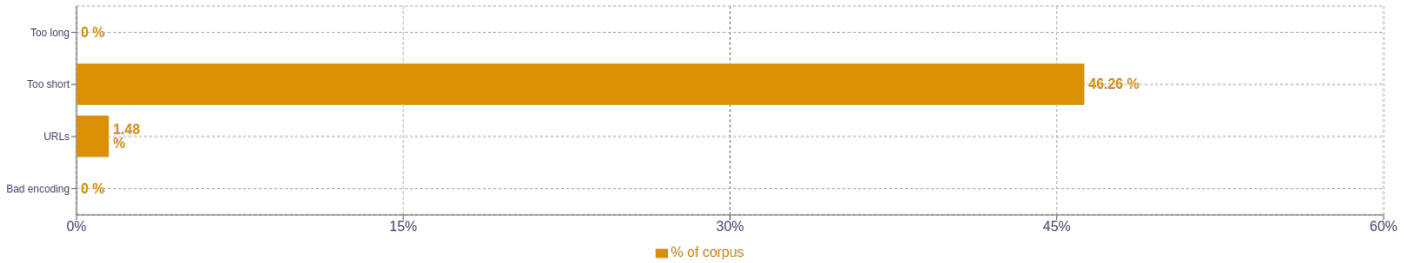


Segment length distribution by token

<= 49 tokens = **5.7M** segments | **26M** duplicates
> 50 tokens = **1.5M** segments | **363K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	અને | 1042879 · to | 867809 · the | 840854 · in | 793484 · news | 786731
2	all rights | 123055 · rights reserved | 122689 · pm ist | 114033 · privacy policy | 109168 · gujarati news | 103074
3	all rights reserved | 122480 · skip to content | 56309 · your email address | 51537 · i am gujarat | 50032 · share to twittershare | 41610
4	share to twittershare to | 41605 · twittershare to facebookshare to | 41039 · to twittershare to facebookshare | 41039 · to facebookshare to pinterest | 41037 · leave a reply cancel | 35191
5	to twittershare to facebookshare to | 41039 · share to twittershare to facebookshare | 41039 · twittershare to facebookshare to pinterest | 41037 · leave a reply cancel reply | 35191 · website in this browser for | 33917

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).

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.

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