# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Analytics date | Language |
|---|---|---|
| te_1.jsonl.tsv | 3/21/2024 | Telugu (te) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Size | Characters |
|---|---|---|---|---|---|
| 415,598 | 51,141,409 | 11,030,905<br>(21.57 %) | 573M | 6.75 GB | |

## Top 10 domains

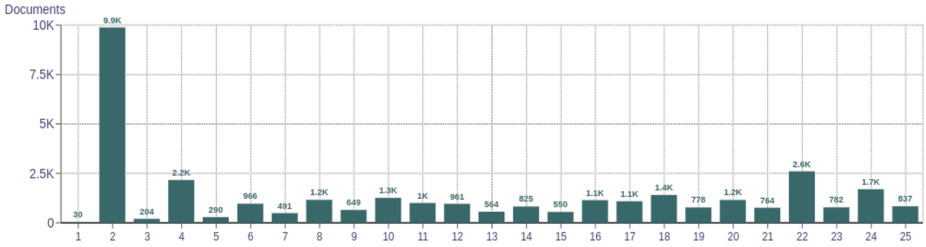
| Domain | Docs | % of total |
|---|---|---|
| blogspot.in | 31K | 7.47 |
| andhrajyothy.com | 27K | 6.47 |
| samayam.com | 15K | 3.53 |
| blogspot.com | 14K | 3.39 |
| newsmeter.in | 11K | 2.75 |
| wikipedia.org | 9.1K | 2.19 |
| eenadu.net | 9K | 2.17 |
| news18.com | 8.9K | 2.15 |
| studysite.org | 7.2K | 1.73 |
| asianetnews.com | 5.9K | 1.41 |

## Top 10 TLDs

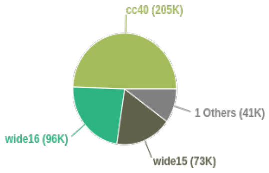
| Domain | Docs | % of total |
|---|---|---|
| com | 257K | 61.83 |
| in | 71K | 17.16 |
| org | 27K | 6.55 |
| net | 24K | 5.67 |
| co.in | 4.9K | 1.17 |
| ae | 4.4K | 1.05 |
| sg | 2.8K | 0.67 |
| page | 2.6K | 0.64 |
| info | 2K | 0.48 |
| pt | 1.4K | 0.34 |

## Documents size (in segments)

**<= 25** segments **8.04%** (33K documents)
**> 25** segments **91.96%** (380K documents)

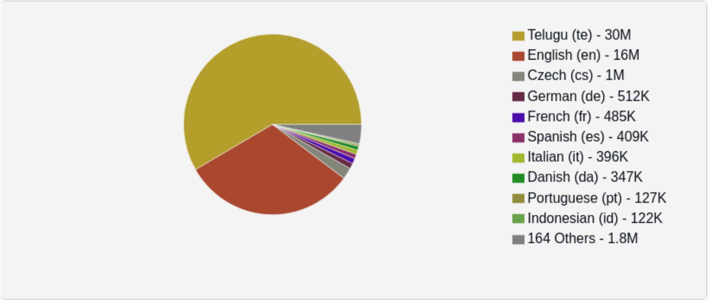


## Documents by collection

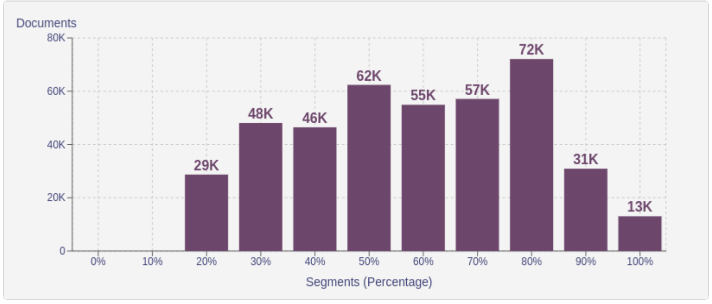


## Language Distribution

### Number of segments

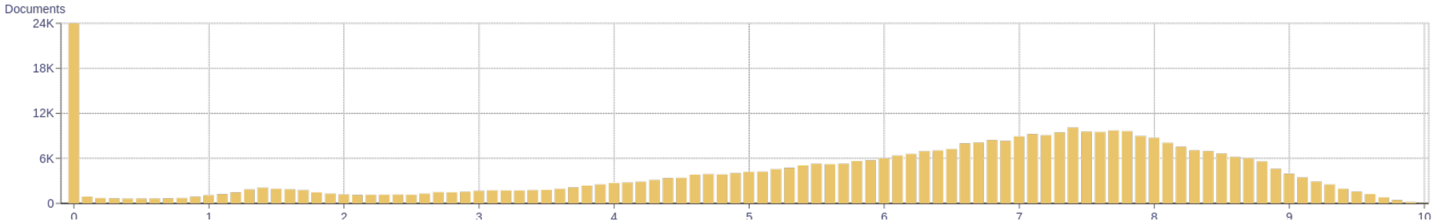


### Percentage of segments in Telugu (te) inside documents



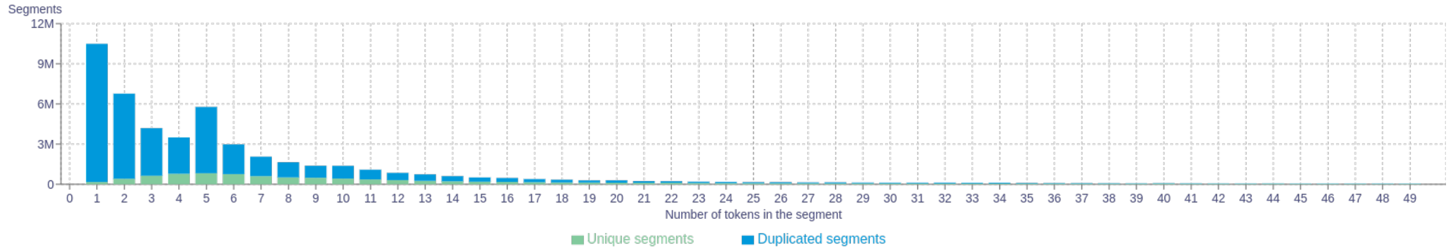
## Distribution of documents by document score

score <= 5 - **27.01%** (112K documents)
score > 5 - **72.99%** (303K documents)

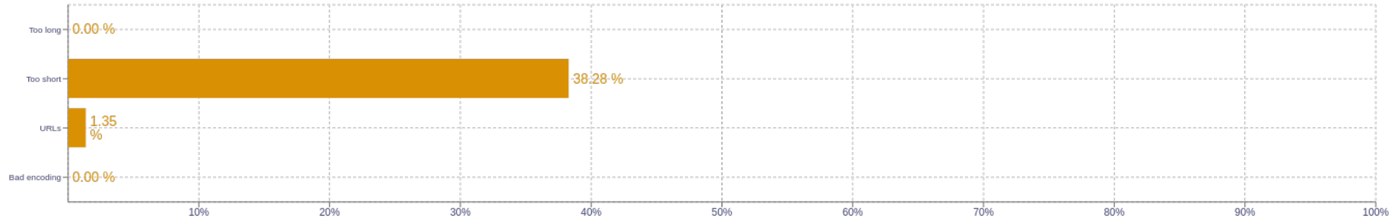


## Segment length distribution by token

**<= 49** tokens = **9.5M** segments | **39M** duplicates
**> 50** tokens = **2.1M** segments | **621K** duplicates



## Segment noise distribution



### Frequent n-grams

| Size | n-grams |
|---|---|
| 1 | ఈ \| 2660949<br>the \| 1952812<br>to \| 1857841<br>news \| 1504450<br>ఆ \| 1483233 |
| 2 | span style \| 216760<br>telugu news \| 216156<br>of the \| 203200<br>posted by \| 198221<br>read more \| 198196 |
| 3 | all rights reserved \| 147922<br>to twittershare to \| 121391<br>share to twittershare \| 121391<br>twittershare to facebookshare \| 118860<br>to facebookshare to \| 118860 |
| 4 | share to twittershare to \| 121391<br>twittershare to facebookshare to \| 118860<br>to twittershare to facebookshare \| 118860<br>to facebookshare to pinterest \| 118860<br>శ్రీరామ శ్రీరామ శ్రీరామ శ్రీరామ \| 95956 |
| 5 | twittershare to facebookshare to pinterest \| 118860<br>to twittershare to facebookshare to \| 118860<br>share to twittershare to facebookshare \| 118860<br>శ్రీరామ శ్రీరామ శ్రీరామ శ్రీరామ శ్రీరామ \| 82331<br>భాగస్వామ్యం చెయ్యండిfacebookకు భాగస్వామ్యం చెయ్యండిpinterestకు భాగస్వామ్యం \| 31802 |

## 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