HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
mai_Deva.jsonl.tsv	12/4/2024	Maithili (mai)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
24,979	645,527	368,607 (57.10 %)	21M	96,119,799	233.26 MB

Top 10 domains

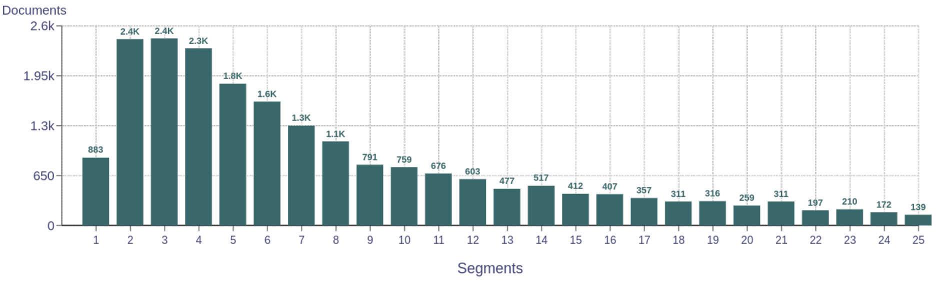
Domain	Docs	% of total
esamaad.com	3.6K	14.30
hellomithila.com	3.3K	13.10
wikipedia.org	2.8K	11.17
mithiladainik.in	2.5K	10.19
blogspot.com	2.5K	10.16
maithilijindabaad.com	2.3K	9.34
mithimedia.in	1.7K	6.91
blogspot.in	936	3.75
mithilamirror.com	745	2.98
mithila.live	408	1.63

Top 10 TLDs

Domain	Docs	% of total
com	14K	57.71
in	6K	24.16
org	3.3K	13.15
live	408	1.63
pl	308	1.23
de	134	0.54
co.in	114	0.46
org.np	110	0.44
com.np	46	0.18
org.in	40	0.16

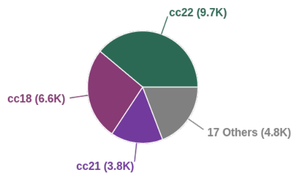
Documents size (in segments)

<= 25 segments **83.35%** (21K documents)
> 25 segments **16.65%** (4.2K documents)



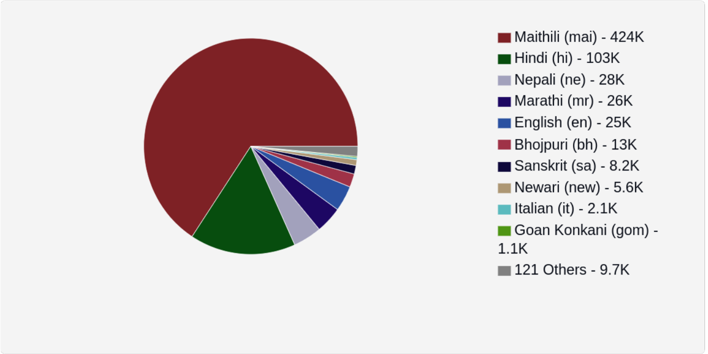
Documents by collection

CC = 85.91%
IA = 14.09%

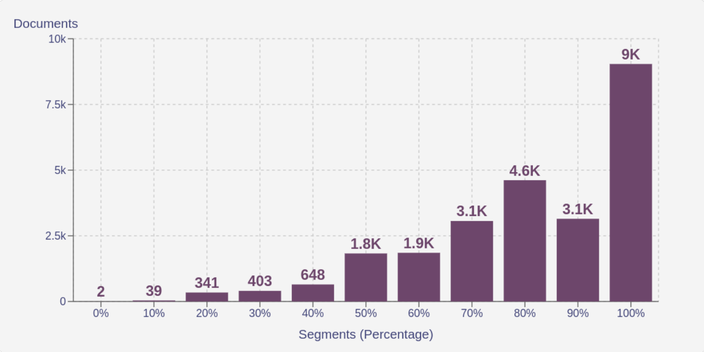


Language Distribution

Number of segments in the Maithili (mai) corpus

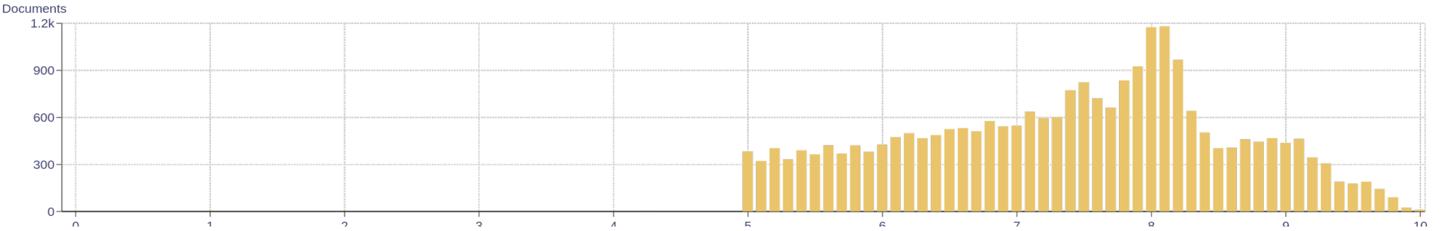


Percentage of segments in Maithili (mai) inside documents



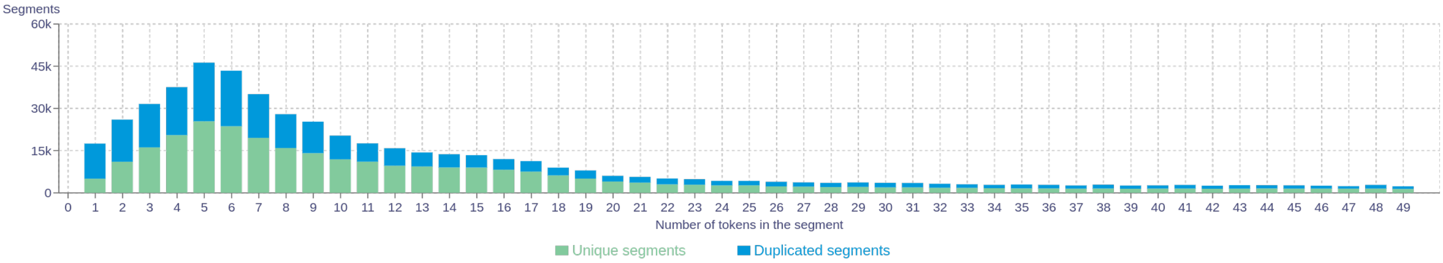
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (25K documents)

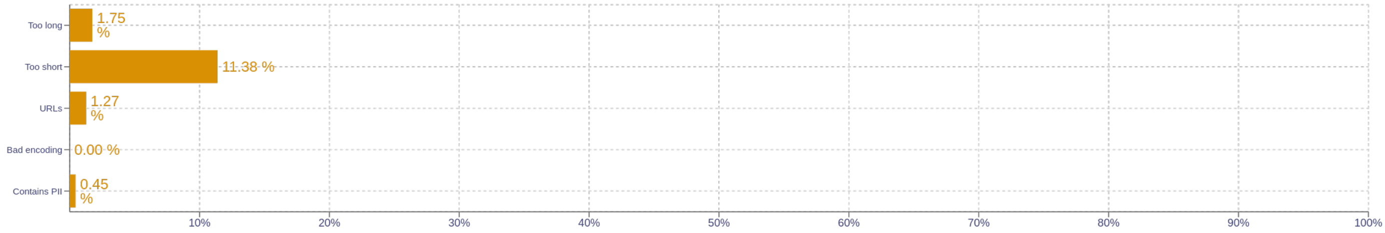


Segment length distribution by token

≤ 49 tokens = **298K** segments | **229K** duplicates
> 50 tokens = **119K** segments | **48K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	अछि | 351838 जे | 169774 सँ | 138410 लेल | 101858 ई | 94005
2	रहल अछि | 40853 अछि जे | 25593 जाइत अछि | 15549 सकैत अछि | 12197 होइत अछि | 11677
3	देल गेल अछि | 3463 कएल गेल अछि | 2927 रहल अछि जे | 1960 मुख्यमंत्री नीतीश कुमार | 1856 कहल जाइत अछि | 1827
4	पुरनका लेख सर्च करु | 1051 सदस्यता लें टिप्पणियाँ भेजें | 1010 देल जा रहल अछि | 884 कैल जा रहल अछि | 812 प्रथम मैथिली पाक्षिक ई | 810
5	प्रथम मैथिली पाक्षिक ई पत्रिका | 601 सङ्घीय मामिला तथा स्थानीय विकास | 498 सालमे सङ्घीय मामिला तथा स्थानीय | 438 मामिला तथा स्थानीय विकास मन्त्रालयद्वारा | 381 प्रसाद मण्डलक दूटा लघु कथा | 263

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