HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
awa_Deva.jsonl.tsv	10/3/2024	Awadhi (awa)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
7,281	131,475	70,186 (53.38 %)	6.9M	67.99 MB	28,649,068

Top 10 domains

Domain	Docs	% of total
biblegateway.com	2.1K	28.53
khabarlahariya.org	1.5K	20.04
bible.is	503	6.91
wikipedia.org	412	5.66
awadh.org	207	2.84
districtsinindia.com	156	2.14
blogspot.com	106	1.46
gospelgo.com	103	1.41
bharatdiscovery.org	78	1.07
blogspot.in	59	0.81

Top 10 TLDs

Domain	Docs	% of total
com	3.9K	54.11
org	2.3K	31.22
is	503	6.91
in	346	4.75
net	72	0.99
co.in	26	0.36
page	23	0.32
gov.in	12	0.16
nic.in	10	0.14
info	10	0.14

Documents size (in segments)

<= 25 segments **80.39%** (5.9K documents)
> 25 segments **19.61%** (1.4K documents)
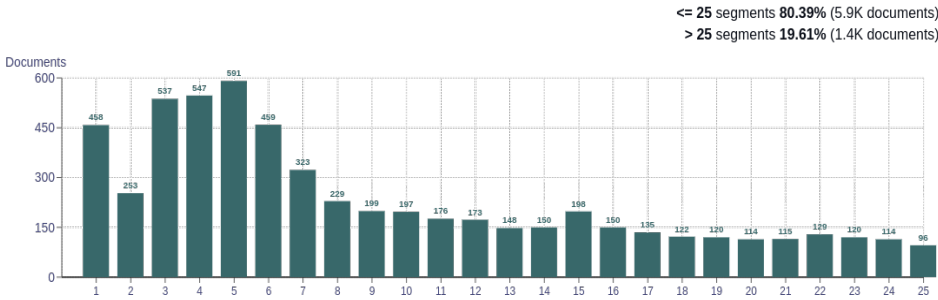
Documents by collection

cc18 (2.2K)
cc22 (1.4K)
cc17 (845)
18 Others (2.9K)
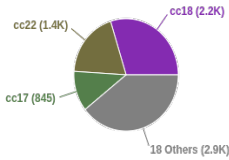
Language Distribution

Number of segments

- Awadhi (awa) - 80K
- Hindi (hi) - 31K
- English (en) - 8.6K
- Marathi (mr) - 6.7K
- Nepali (ne) - 1.6K
- Sanskrit (sa) - 462
- Newari (new) - 442
- Italian (it) - 287
- French (fr) - 262
- Bhojpuri (bh) - 231
- 70 Others - 1.3K

*Awadhi (awa) identification might be inaccurate because language is not supported by Fasttext
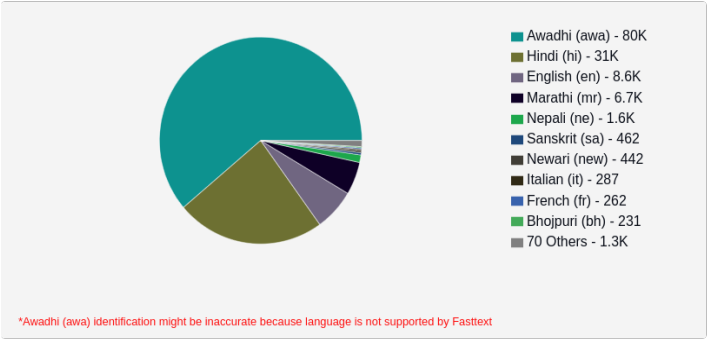
Percentage of segments in Awadhi (awa) inside documents
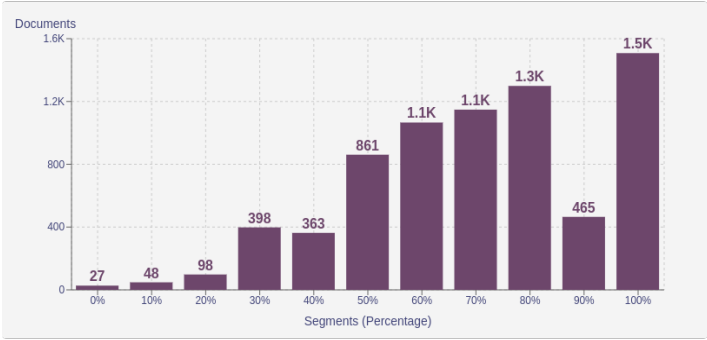
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (7.3K documents)

Segment length distribution by token

<= 49 tokens = **54K** segments | **46K** duplicates
> 50 tokens = **31K** segments | **15K** duplicates

- Unique segments
- Duplicated segments

Segment noise distribution

Too long	0.00 %
Too short	5.93 %
URLs	0.29 %
Bad encoding	0.00 %
Contains PII	0.02 %
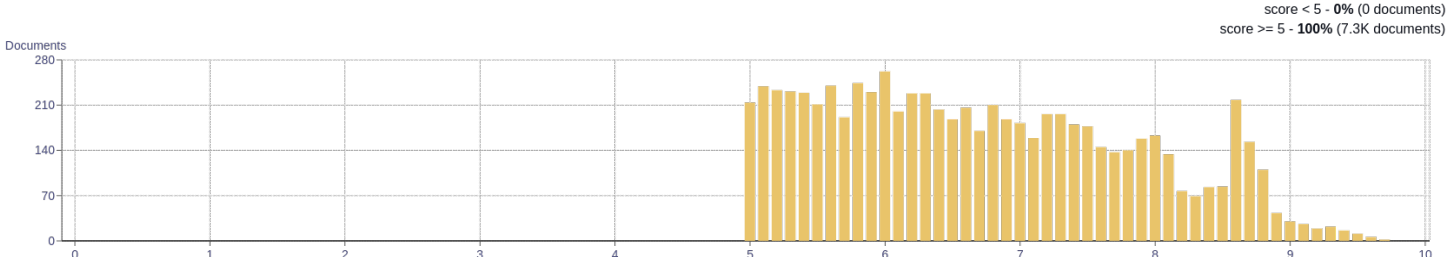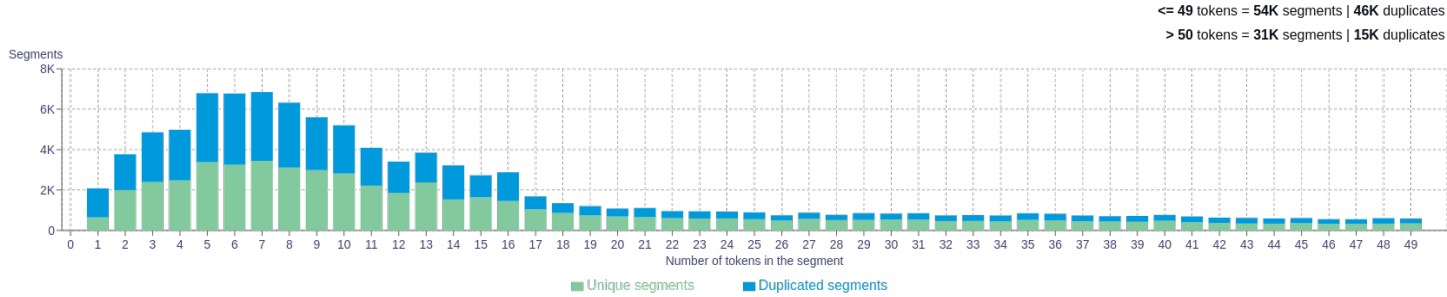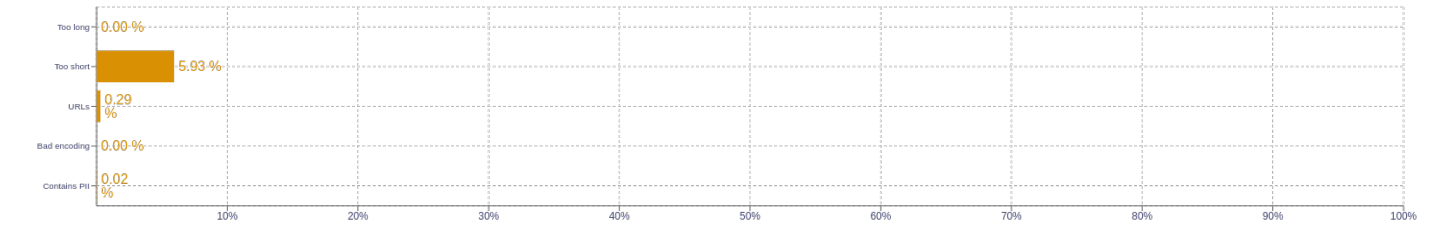
Frequent n-grams

Size	n-grams
1	यहोवा | 28061 और | 13950 ईसू | 13775 ओकरे | 13426 जात | 10588
2	read version | 3366 उत्तर प्रदेश | 2747 awadhi bible | 2419 चाहिँ कईंती | 1484 in hindi | 1235
3	world bible translation | 648 read version copyright | 648 bible translation center | 648 यहोवा क समन्या | 582 इसाएल क मनइयन | 548
4	world bible translation center | 648 ने मनाया सुहाग रात | 215 और अंजना ने मनाया | 215 अंजना ने मनाया सुहाग | 215 निरहुआ और अंजना ने | 214
5	और अंजना ने मनाया सुहाग | 215 अंजना ने मनाया सुहाग रात | 215 निरहुआ और अंजना ने मनाया | 214 यहोवा तोहार परमेस्सर तू पचनक | 195 nirahua anjana singh bhojpuri film | 192

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