

General overview

| Corpus | Analytics date | Language |
|--------------------|----------------|--------------|
| awa_Deja.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 |

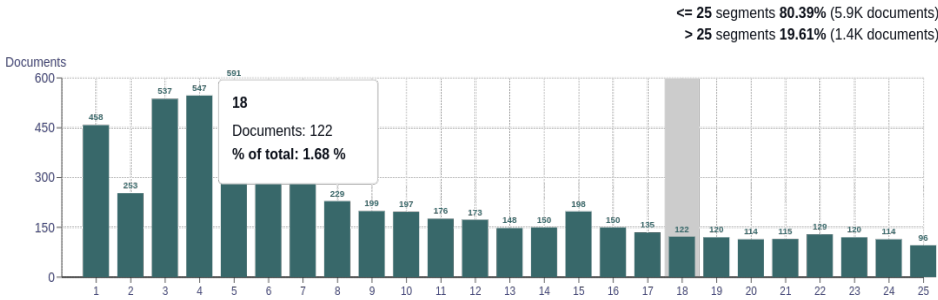
Top 10 domains

| 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 |

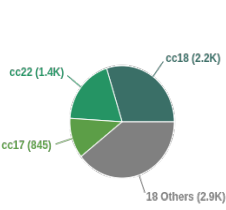
Top 10 TLDs

| 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)

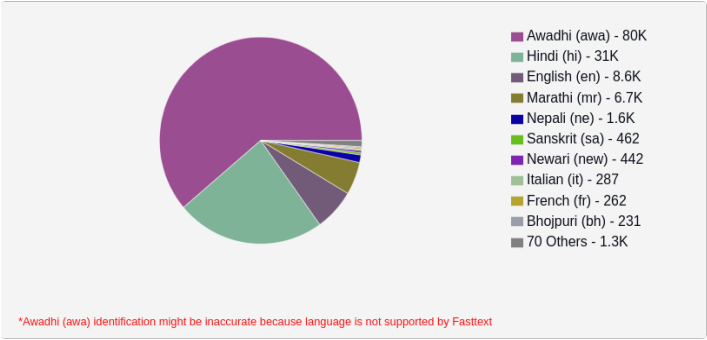


Documents by collection

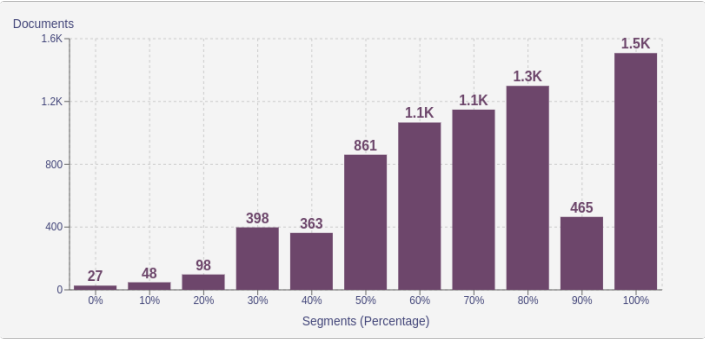


Language Distribution

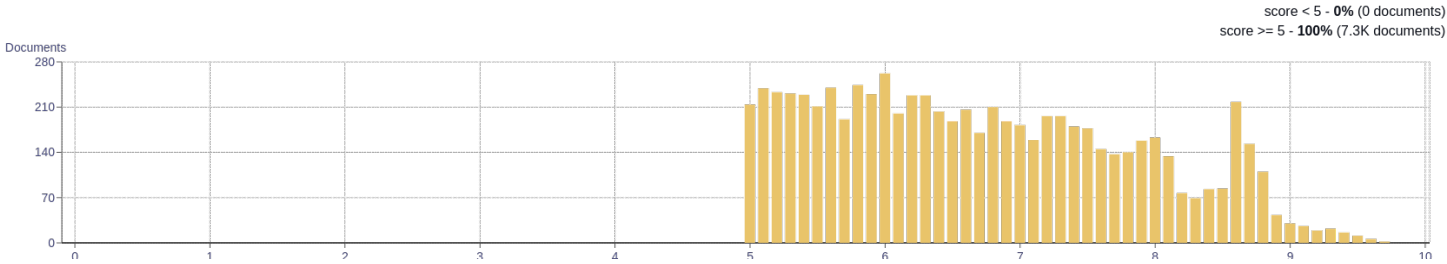
Number of segments



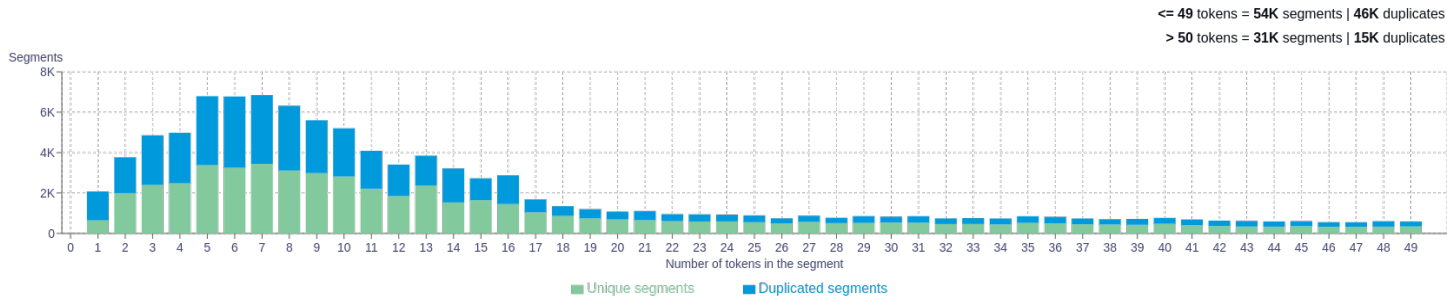
Percentage of segments in Awadhi (awa) inside documents



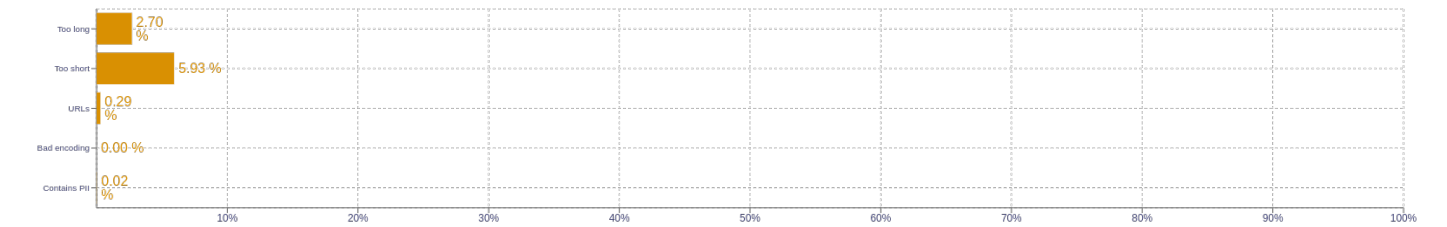
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

| Size | n-grams |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1 | <div>यहोवा 28061</div> <div>और 13950</div> <div>ईसू 13775</div> <div>औकने 13426</div> <div>जात 10588</div> |
| 2 | <div>read version 3366</div> <div>उत्तर प्रदेश 2747</div> <div>awadhi bible 2419</div> <div>चारिहुँ कहली 1484</div> <div>in hindi 1235</div> |
| 3 | <div>world bible translation 648</div> <div>read version copyright 648</div> <div>bible translation center 648</div> <div>यहोवा क समन्या 582</div> <div>इस्राएल क मन्हुयन 548</div> |
| 4 | <div>world bible translation center 648</div> <div>ने मनाया सुहाग रात 215</div> <div>और अंजना ने मनाया 215</div> <div>अंजना ने मनाया सुहाग 215</div> <div>निरहुआ और अंजना ने 214</div> |
| 5 | <div>और अंजना ने मनाया सुहाग 215</div> <div>अंजना ने मनाया सुहाग रात 215</div> <div>निरहुआ और अंजना ने मनाया 214</div> <div>यहोवा तोहार परमेस्सर तू पक्क 195</div> <div>nirahua anjana singh bhojpuri film 192</div> |

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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>