

General overview

Corpus	Analytics date	Language
HPLT-docslite.ur.tsv	6/8/2024	Urdu (ur)

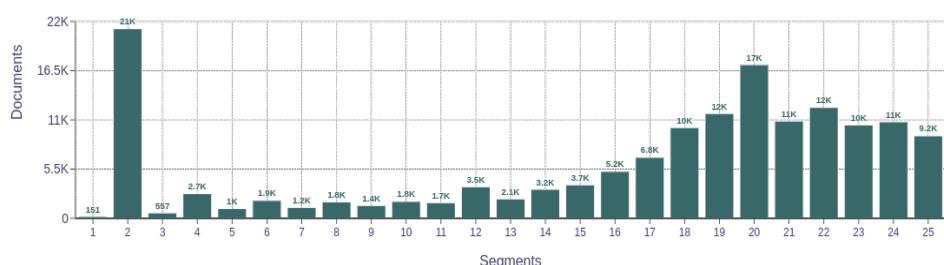
Volumes

Docs	Segments	Unique segments	Tokens	Size
1,437,830	136,409,453	55,026 (0.04 %)	1.7B	10.61 GB

Type-Token Ratio

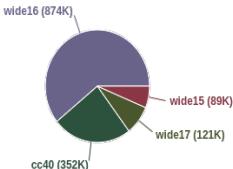
Urdu (ur)

Documents size (in segments)



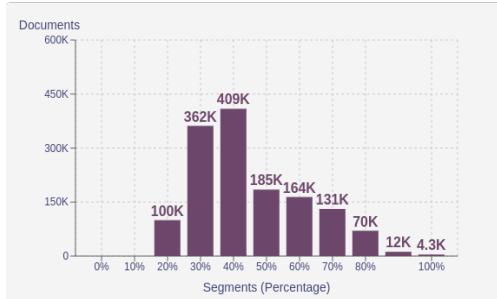
<= 25 segments 10.59% (152K documents)
> 25 segments 89.41% (1.3M documents)

Documents by collection

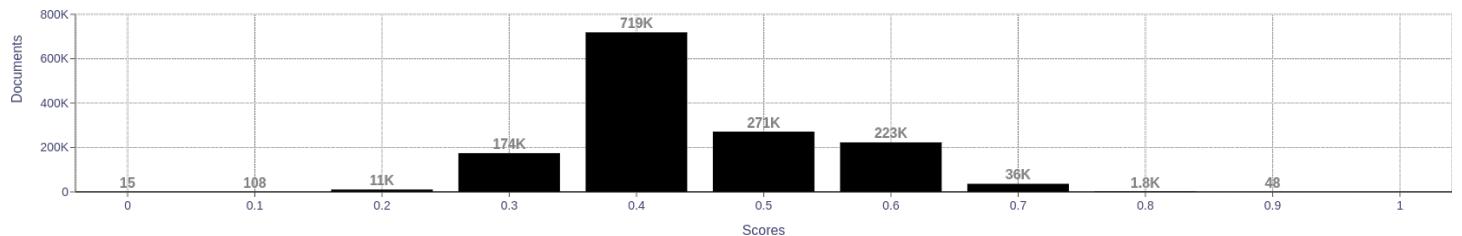


Language Distribution

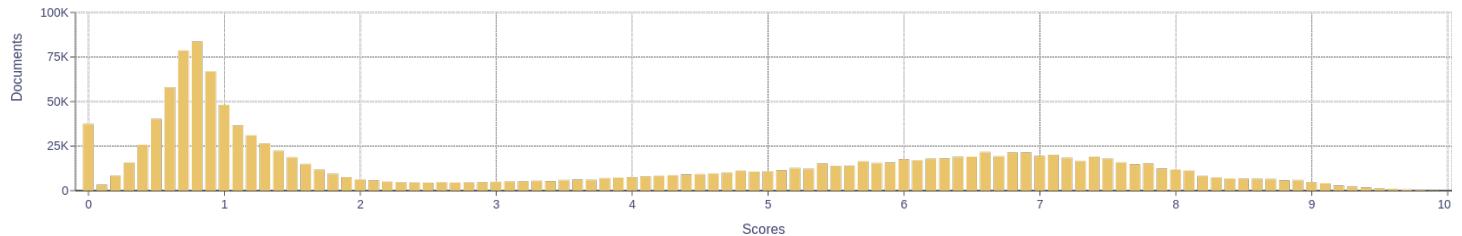
Percentage of segments in Urdu (ur) inside documents



Distribution of documents by average fluency score



Distribution of documents by document score



Segment length distribution by token



<= 49 tokens = 28M segments | 103M duplicates
> 50 tokens = 5.1M segments | 1.1M duplicates

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/mabanon/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>