HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
nus_Latn.jsonl.tsv	11/5/2024	Nuer (nus)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
272	8,514	2,720 (31.95 %)	583K	2.13 MB	1,873,800

Top 10 domains

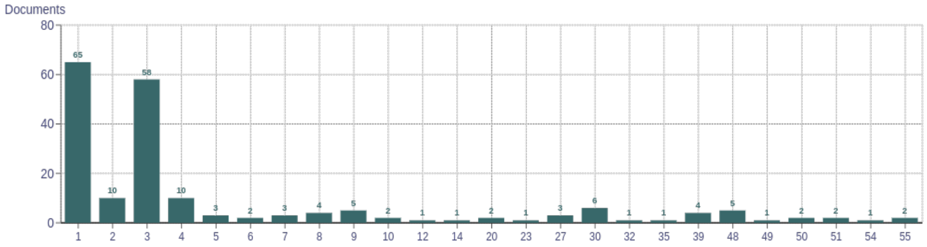
Domain	Docs	% of total
bible.is	115	42.28
indiana.edu	65	23.90
consumer.vic.gov.au	27	9.93
jw.org	25	9.19
nanetya-foundation.org	9	3.31
ironchariots.org	7	2.57
triquidechicahuaxtla.org	5	1.84
multicultural.vic.gov.au	3	1.10
4laws.com	3	1.10
nueronline.com	2	0.74

Top 10 TLDs

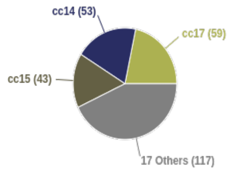
Domain	Docs	% of total
is	115	42.28
edu	65	23.90
org	47	17.28
vic.gov.au	33	12.13
com	9	3.31
com.au	3	1.10

Documents size (in segments)

<= 25 segments **71.69%** (195 documents)
> 25 segments **28.31%** (76 documents)

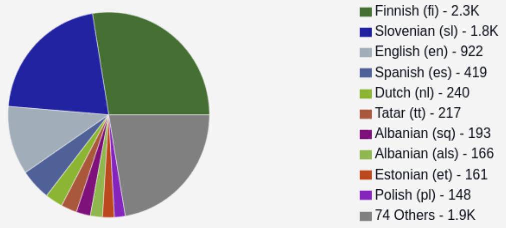


Documents by collection



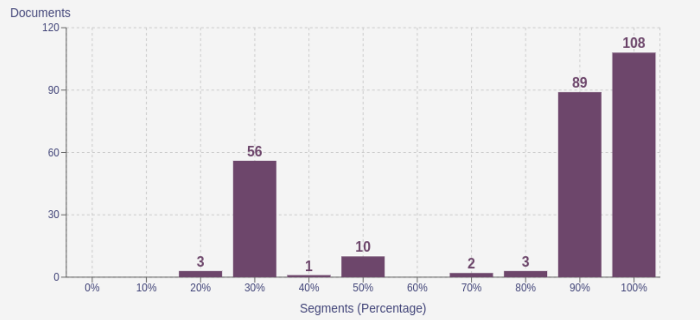
Language Distribution

Number of segments



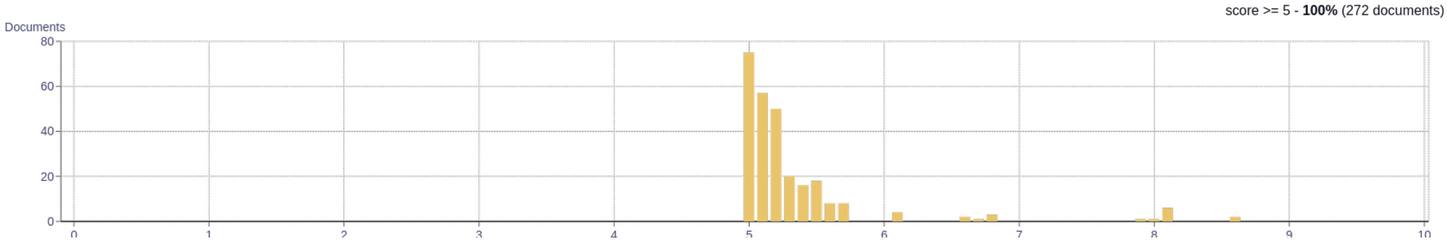
*Nuer (nus) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Nuer (nus) inside documents



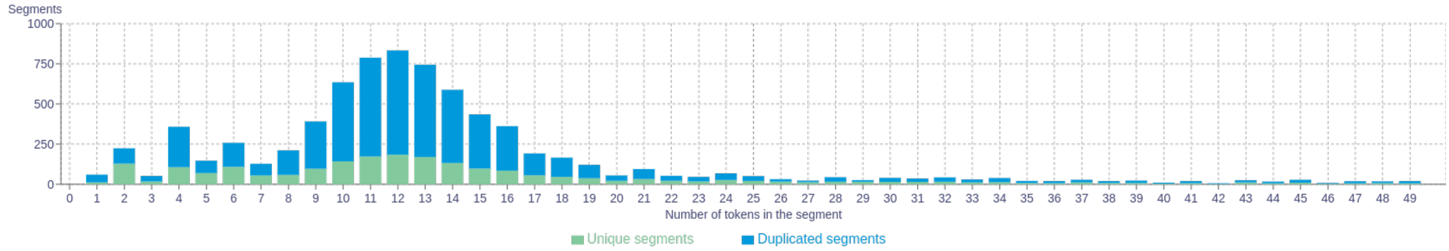
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (272 documents)

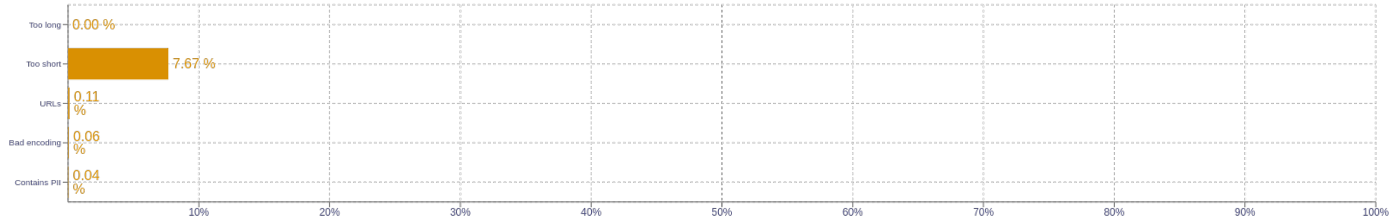


Segment length distribution by token

<= 49 tokens = **2.2K** segments | **5.5K** duplicates
> 50 tokens = **857** segments | **301** duplicates



Segment noise distribution



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