HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
kas_Arab.jsonl.tsv	12/3/2024	Kashmiri (ks)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
949	27,108	11,770 (43.42 %)	708K	5.87 MB	3,441,826

Top 10 domains

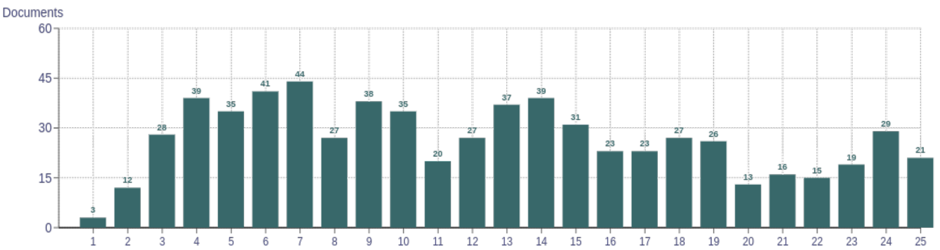
Domain	Docs	% of total
muneeburrahman.com	302	31.82
neabmagazine.com	254	26.77
wikipedia.org	157	16.54
neabinternational.org	52	5.48
wiktionary.org	45	4.74
newschecker.in	29	3.06
gotquestions.org	24	2.53
vikaspedia.in	19	2.00
aminkamil.blog	17	1.79
koshurakhbar.com	13	1.37

Top 10 TLDs

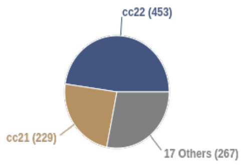
Domain	Docs	% of total
com	594	62.59
org	280	29.50
in	50	5.27
blog	17	1.79
ir	3	0.32
net	2	0.21
ae	1	0.11
ru	1	0.11
io	1	0.11

Documents size (in segments)

<= 25 segments **70.39%** (668 documents)
> 25 segments **29.61%** (281 documents)

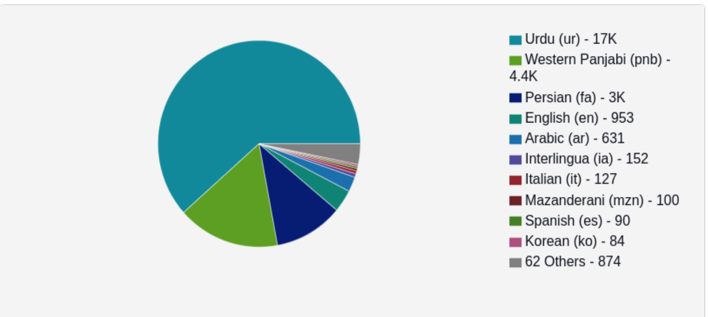


Documents by collection



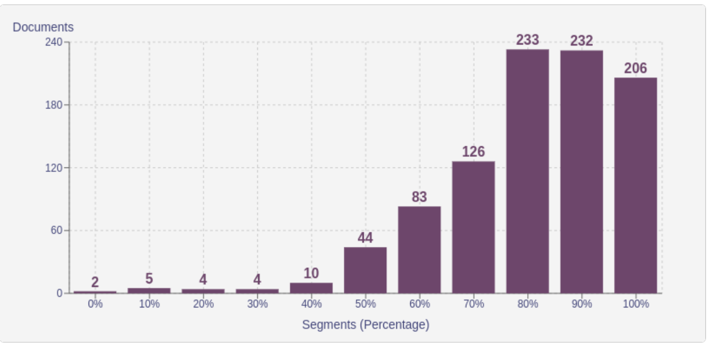
Language Distribution

Number of segments



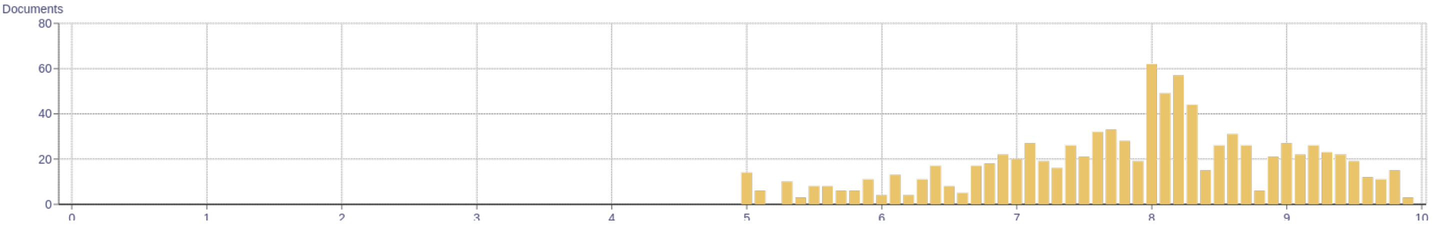
*Kashmiri (ks) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Kashmiri (ks) inside documents



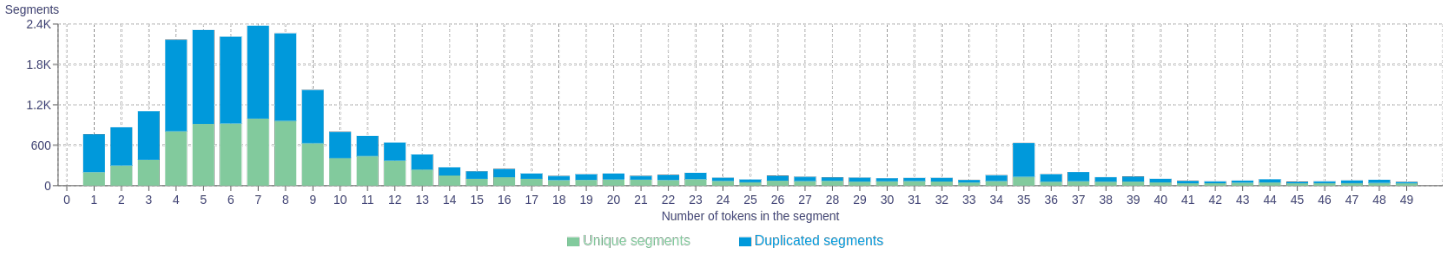
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (949 documents)

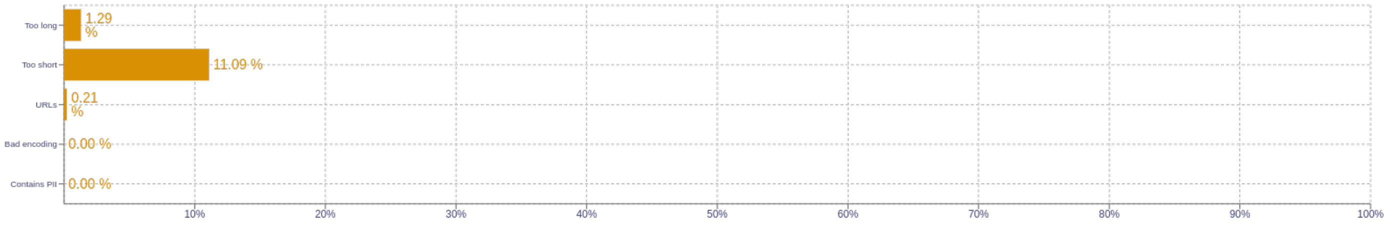


Segment length distribution by token

<= 49 tokens = **10K** segments | **13K** duplicates
> 50 tokens = **3.7K** segments | **2K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	14742 | تہ 14677 | چھ 12340 | منز 7439 | یہ 7036 | نہ
2	5864 | بنر فلکس 5842 | فلکس بنر 1824 | چھ نہ 1306 | منز چھ 699 | چھیہ نہ
3	5840 | بنر فلکس بنر 5838 | فلکس بنر فلکس 235 | اَوَلی ہند یؤرَپی 194 | منز چھ نہ 176 | تم تم تم
4	5836 | بنر فلکس بنر فلکس 5828 | فلکس بنر فلکس بنر 173 | تم تم تم تم 127 | ہند یؤرَپی پیٹھ آمِتي 121 | اَوَلی ہند یؤرَپی پیٹھ
5	5826 | بنر فلکس بنر فلکس بنر 5824 | فلکس بنر فلکس بنر فلکس 171 | تم تم تم تم تم 121 | اَوَلی ہند یؤرَپی پیٹھ آمِتي 65 | کرنس منز حدہ رۄس خوشیہ

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