HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
pa_1.jsonl.tsv	3/17/2024	Punjabi (pa)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
152,775	17,180,972	3,877,316 (22.57 %)	219M	2.03 GB	

Top 10 domains

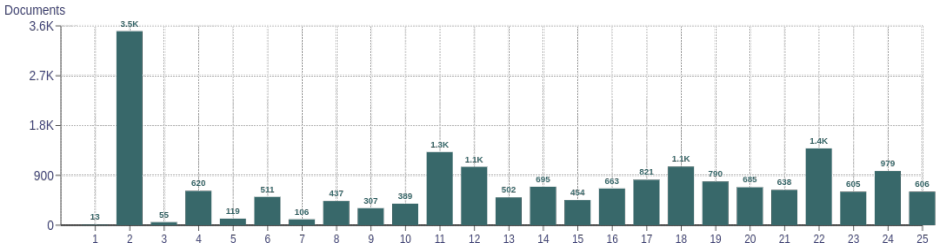
Domain	Docs	% of total
pornk-org.com	6.4K	4.21
news18.com	6.1K	3.96
punjabkesari.in	4.6K	2.99
truescoopnews.com	4.3K	2.82
wikipedia.org	4K	2.60
khalsanews.org	3.9K	2.55
ajitjalandhar.com	3.6K	2.36
quamiekta.com	3.5K	2.26
sikhsiyasat.info	2.6K	1.72
blogspot.in	2.5K	1.64

Top 10 TLDs

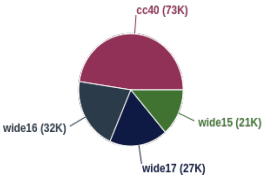
Domain	Docs	% of total
com	97K	63.22
in	19K	12.29
org	16K	10.27
ca	4.5K	2.94
info	3.3K	2.13
co.in	2.6K	1.71
net	1.7K	1.12
com.au	928	0.61
news	845	0.55
xyz	813	0.53

Documents size (in segments)

<= 25 segments **11.99%** (18K documents)
> 25 segments **88.01%** (134K documents)

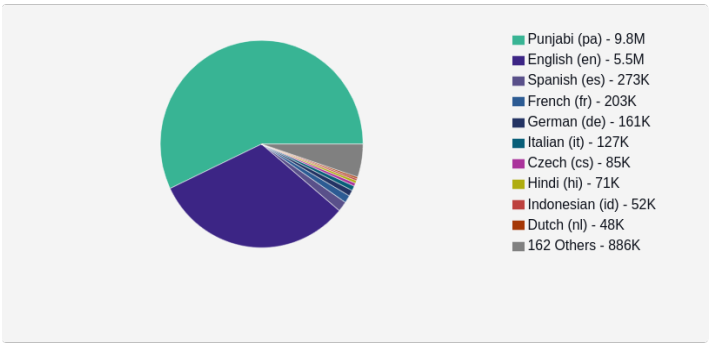


Documents by collection

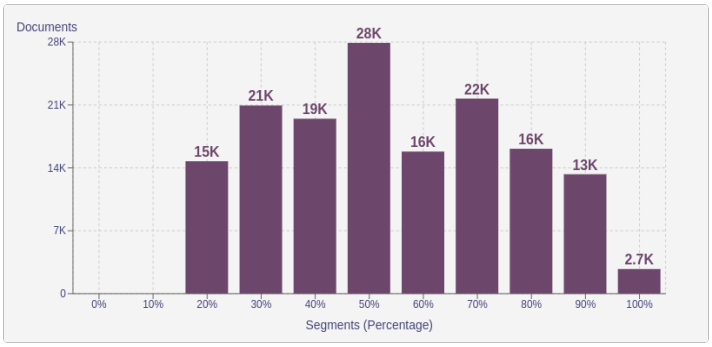


Language Distribution

Number of segments

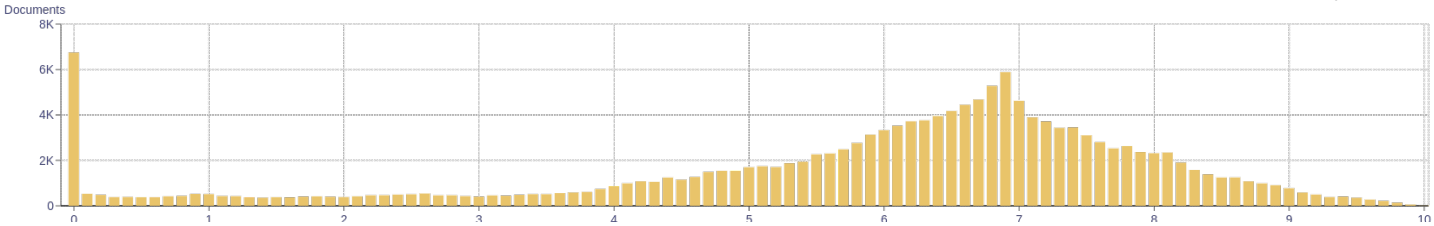


Percentage of segments in Punjabi (pa) inside documents



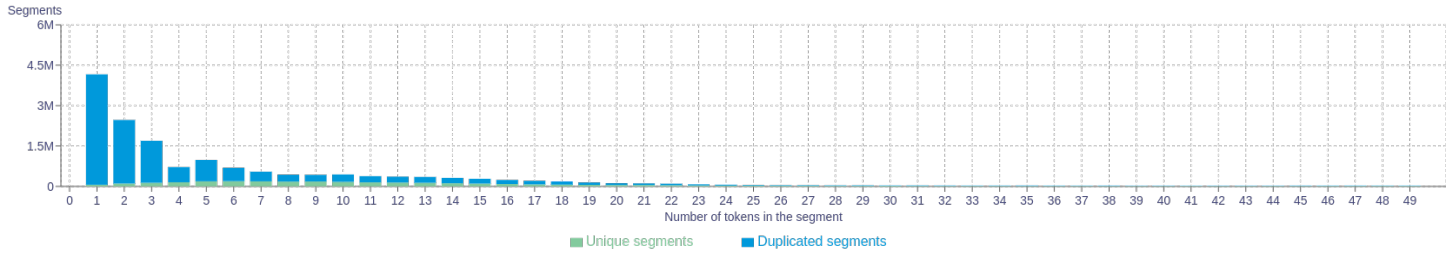
Distribution of documents by document score

score < 5 - **24.13%** (37K documents)
score >= 5 - **75.87%** (116K documents)

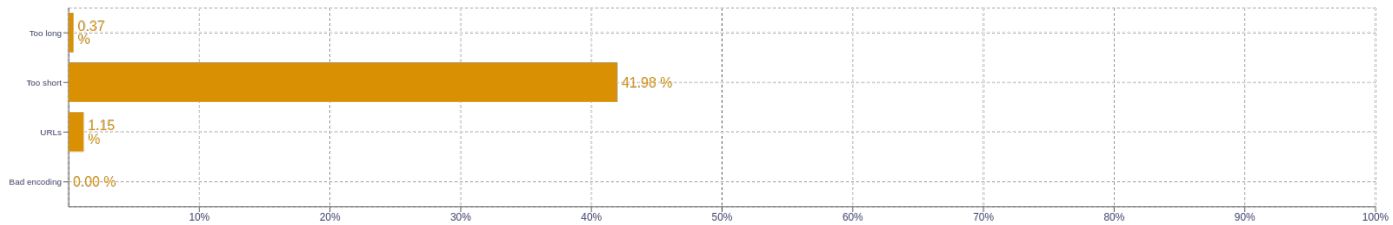


Segment length distribution by token

<= 49 tokens = **3.2M** segments | **13M** duplicates
> 50 tokens = **903K** segments | **211K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	the | 770337 · to | 535359 · of | 442946 · and | 424906 · news | 417666
2	of the | 101145 · all rights | 70140 · rights reserved | 69793 · contact us | 67417 · read more | 61786
3	all rights reserved | 69680 · to twittershare to | 27188 · share to twittershare | 27188 · twittershare to facebookshare | 26828 · to facebookshare to | 26828
4	share to twittershare to | 27188 · twittershare to facebookshare to | 26828 · to twittershare to facebookshare | 26828 · to facebookshare to pinterest | 26828 · leave a reply cancel | 19968
5	twittershare to facebookshare to pinterest | 26828 · to twittershare to facebookshare to | 26828 · share to twittershare to facebookshare | 26828 · leave a reply cancel reply | 19949 · your email address will not | 16384

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