HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
pan_Guru.jsonl.tsv	9/25/2024	Punjabi (pa)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
584,594	11,743,514	6,588,980 (56.11 %)	423M	4.42 GB	1,891,338,690

Top 10 domains

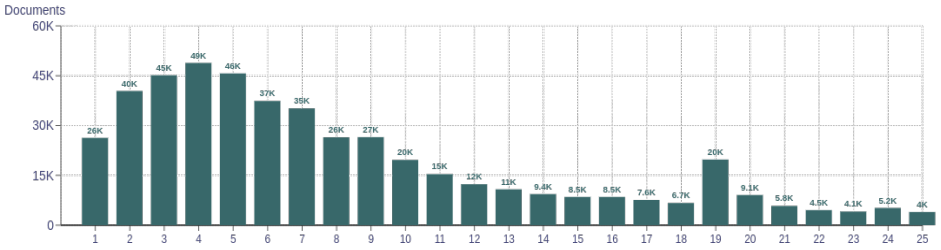
Domain	Docs	% of total
wikipedia.org	28K	4.73
news18.com	21K	3.67
punjabkesari.in	19K	3.22
ajitjalandhar.com	14K	2.39
quamiekta.com	13K	2.17
pornk-org.com	12K	2.12
punjabitribuneonline.com	11K	1.86
dailypost.in	9.1K	1.55
ptcpunjabi.co.in	8.2K	1.41
punjabmailusa.com	8K	1.37

Top 10 TLDs

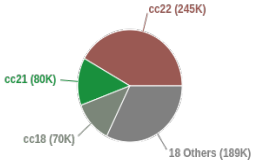
Domain	Docs	% of total
com	376K	64.34
in	67K	11.42
org	62K	10.64
ca	15K	2.55
net	11K	1.86
co.in	9.2K	1.57
info	6.2K	1.07
tv	5.9K	1.00
mobi	3.8K	0.65
news	2.4K	0.41

Documents size (in segments)

<= 25 segments **82.68%** (483K documents)
> 25 segments **17.32%** (101K documents)

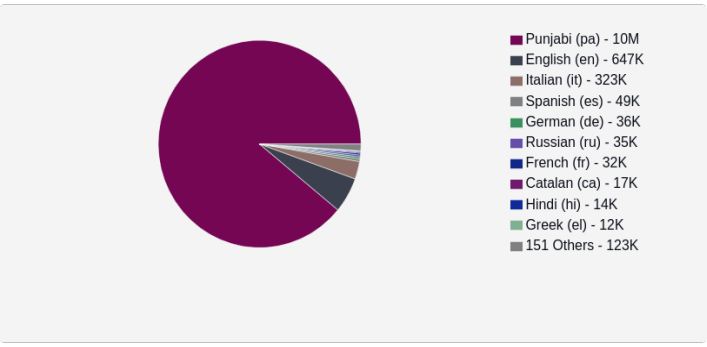


Documents by collection

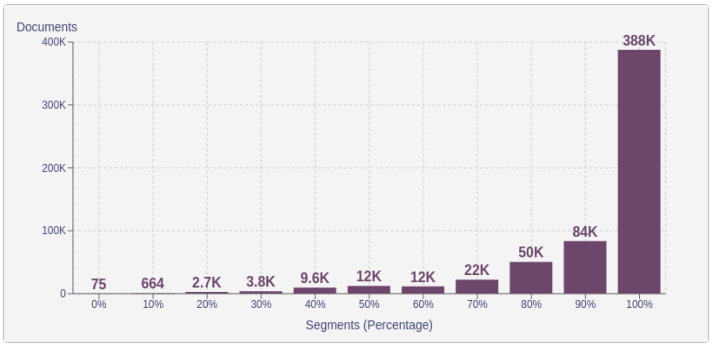


Language Distribution

Number of segments

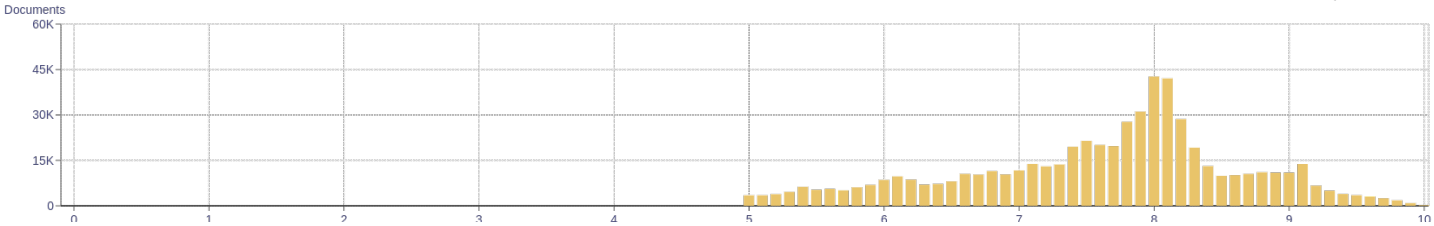


Percentage of segments in Punjabi (pa) inside documents



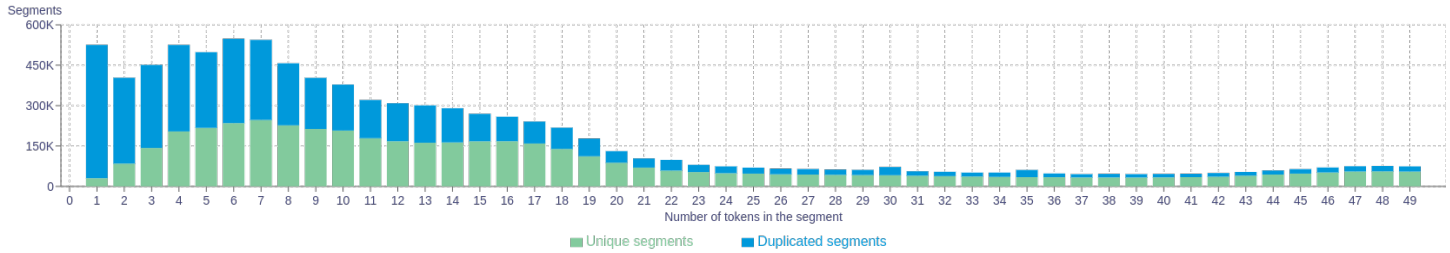
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (585K documents)

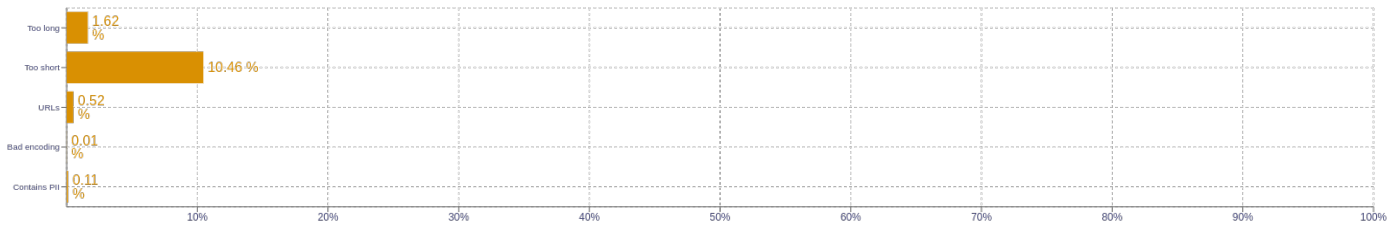


Segment length distribution by token

<= 49 tokens = **4.6M** segments | **4.5M** duplicates
> 50 tokens = **2.7M** segments | **652K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ਮੈਂ | 605499 · ਹਾਂ | 566357 · ‘ਚ | 521965 · ਤੁਸੀਂ | 501798 · ‘ਤੇ | 494570
2	ਪੋਰਨ ਵੀਡੀਓ | 143816 · ਅਕਾਲੀ ਦਲ | 118031 · ਕੈਪਟਨ ਅਮਰਿੰਦਰ | 52389 · ਸ਼ੁਰੋਮਣੀ ਅਕਾਲੀ | 51326 · hours ago | 50417
3	ਸ਼ੁਰੋਮਣੀ ਅਕਾਲੀ ਦਲ | 50443 · sadhu singh hamdard | 39428 · singh hamdard trust | 39427 · ਗੁਰਦੁਆਰਾ ਪ੍ਰਬੰਧਕ ਕਮੇਟੀ | 33312 · ਵਿਧਾਨ ਸਭਾ ਚੋਣਾਂ | 22442
4	sadhu singh hamdard trust | 39427 · ਸ਼ੁਰੋਮਣੀ ਗੁਰਦੁਆਰਾ ਪ੍ਰਬੰਧਕ ਕਮੇਟੀ | 19326 · written consent of the | 13146 · without the prior written | 13146 · whole or in part | 13146
5	written consent of the trust | 13146 · without the prior written consent | 13146 · whole or in part be | 13146 · trust may not in whole | 13146 · to the trust may not | 13146

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