HPLT Analytics report

HPLT Analytics

General overview

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

Volumes

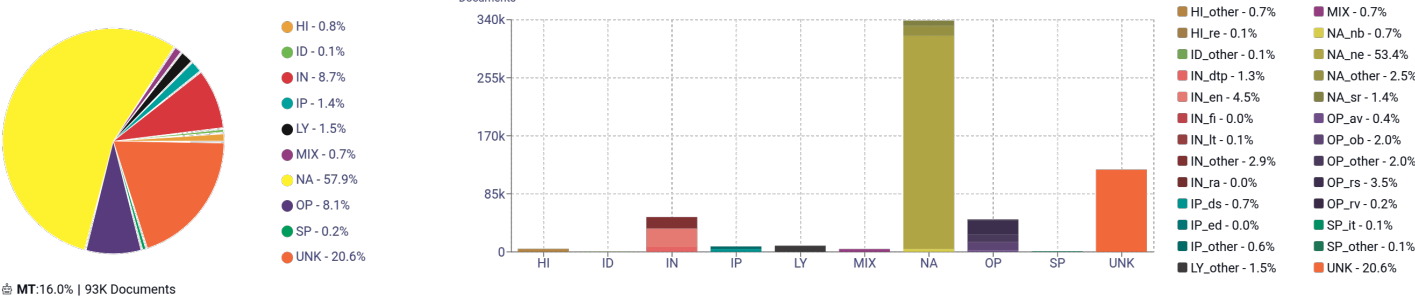| Docs	Segments	Unique segments	Tokens	Characters	Size
584,594	11,743,514	6,588,980 (56.11 %)	423M	1,891,338,690	4.42 GB

Top 10 domains

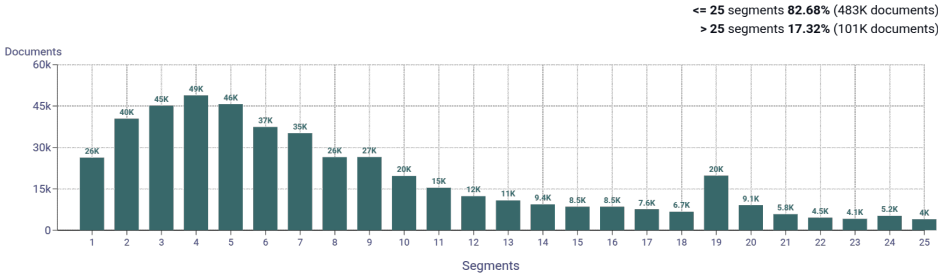| Domain	Docs	% of total
wikipedia.org	28K	4.73%
news18.com	21K	3.67%
punjabkesari.in	19K	3.22%
ajitjalandhar.com	14K	2.39%
quamiekta.com	13K	2.17%
pornk-org.com	12K	2.12%
punjabitribuneo...	11K	1.86%
dailypost.in	9.1K	1.55%
ptcpunjabi.co.in	8.2K	1.41%
punjabmailusa.com	8K	1.37%

Top 10 TLDs

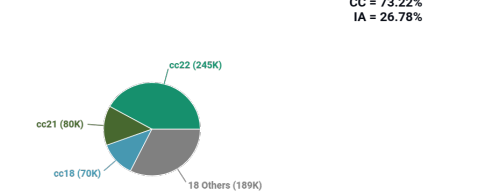| Domain	Docs	% of total
com	376K	64.34%
in	67K	11.42%
org	62K	10.64%
ca	15K	2.55%
net	11K	1.86%
co.in	9.2K	1.57%
info	6.2K	1.07%
tv	5.9K	1.00%
mobi	3.8K	0.65%
news	2.4K	0.41%

Register labels

Pie chart:
- HI - 0.8%
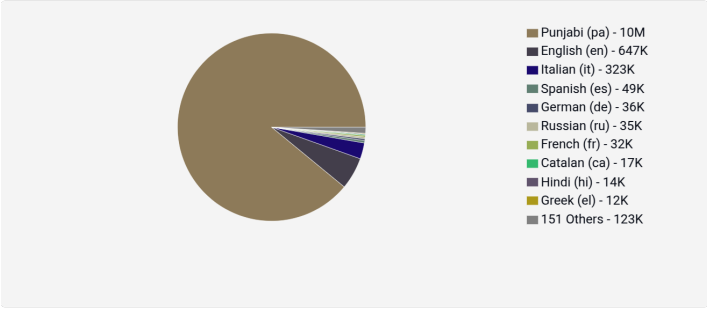- ID - 0.1%
- IN - 8.7%
- IP - 1.4%
- LY - 1.5%
- MIX - 0.7%
- NA - 57.9%
- OP - 8.1%
- SP - 0.2%
- UNK - 20.6%

MT:16.0% | 93K Documents

Bar chart legend:
- HI_other - 0.7%
- HI_re - 0.1%
- ID_other - 0.1%
- IN_dtp - 1.3%
- IN_en - 4.5%
- IN_fi - 0.0%
- IN_lt - 0.1%
- IN_other - 2.9%
- IN_ra - 0.0%
- IP_ds - 0.7%
- IP_ed - 0.0%
- IP_other - 0.6%
- LY_other - 1.5%
- MIX - 0.7%
- NA_nb - 0.7%
- NA_ne - 53.4%
- NA_other - 2.5%
- NA_sr - 1.4%
- OP_av - 0.4%
- OP_ob - 2.0%
- OP_other - 2.0%
- OP_rs - 3.5%
- OP_rv - 0.2%
- SP_it - 0.1%
- SP_other - 0.1%
- UNK - 20.6%

Documents size (in segments)

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

Documents by collection

CC = 73.22%
IA = 26.78%

- cc22 (245K)
- cc21 (80K)
- cc18 (70K)
- 18 Others (189K)

Language Distribution

Number of segments in the Punjabi (pa) corpus

- Punjabi (pa) - 10M
- English (en) - 647K
- Italian (it) - 323K
- Spanish (es) - 49K
- German (de) - 36K
- Russian (ru) - 35K
- French (fr) - 32K
- Catalan (ca) - 17K
- Hindi (hi) - 14K
- Greek (el) - 12K
- 151 Others - 123K

Percentage of segments in Punjabi (pa) inside documents

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (585K documents)
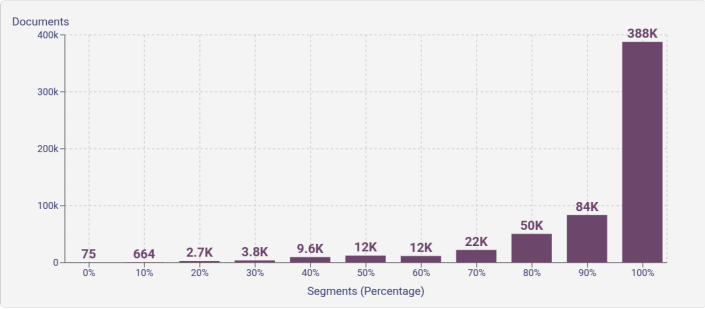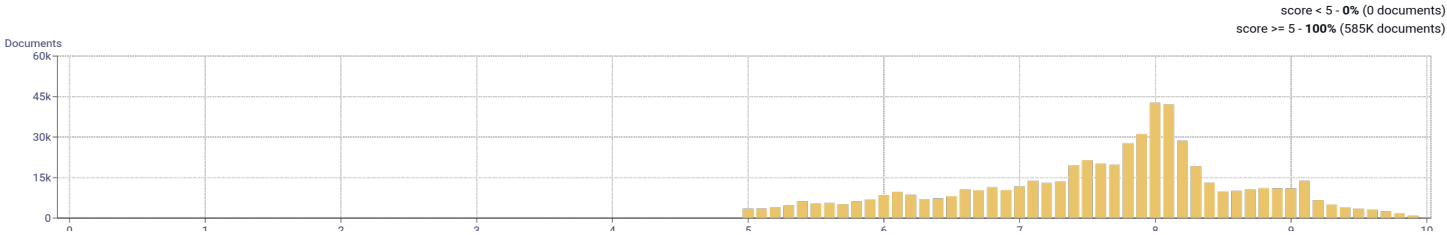
Segment length distribution by token

Segments

Number of tokens in the segment

■ Unique segments ■ Duplicated segments

Segment noise distribution

Too long	1.62 %
Too short	10.46 %
URLs	0.52 %
Bad encoding	0.01 %
Contains PII	0.11 %
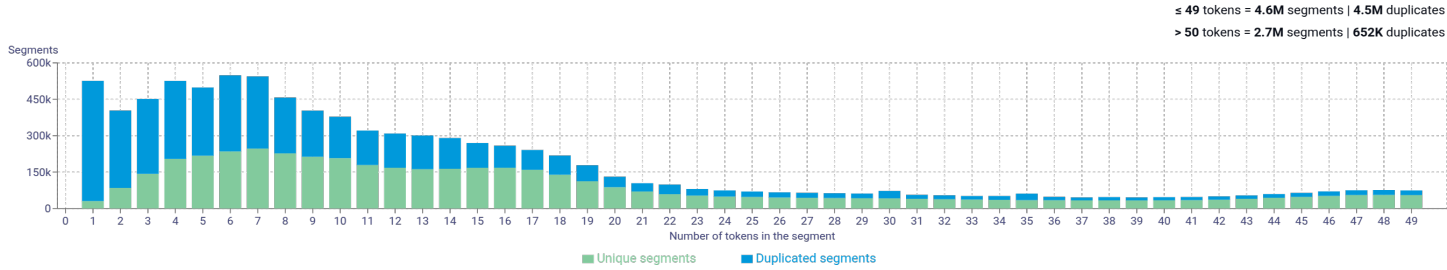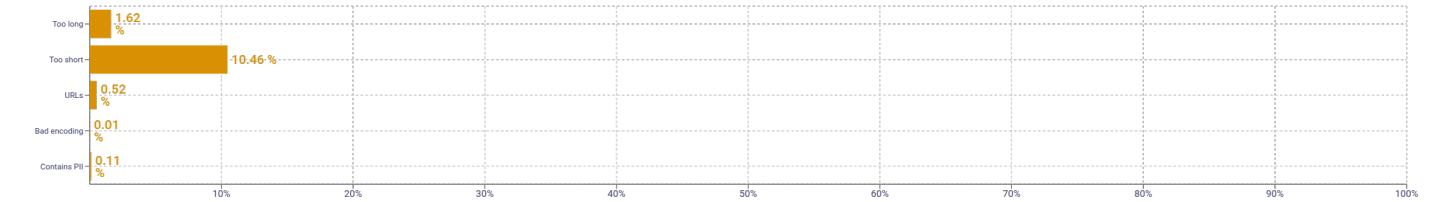
Frequent n-grams

Size	n-grams
1	ਸੈ | 605499 ਹਾਂ | 566357 'ਚ | 521965 ਤੁਸੀਂ | 501798 'ਤੇ | 494570
2	ਘੇਰਨ ਵੀਡੀਓ | 143816 ਅਕਾਲੀ ਦਲ | 118031 ਕੈਪਟਨ ਅਮਰਿੰਦਰ | 52389 ਸ਼੍ਰੋਮਣੀ ਅਕਾਲੀ | 51326 hours ago | 50417
3	ਸ਼੍ਰੋਮਣੀ 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 ਸ਼੍ਰੋਮਣੀ 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

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	How-to or instructions	HI	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	Informational persuasion	IP	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	Opinion	OP
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	Informational description	IN	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				