HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-urd_Arab	9/18/2025	Urdu (ur)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
7,206,563	96,028,325	67,605,412 (70.40 %)	4.4B	19,156,637,140	31.57 GB

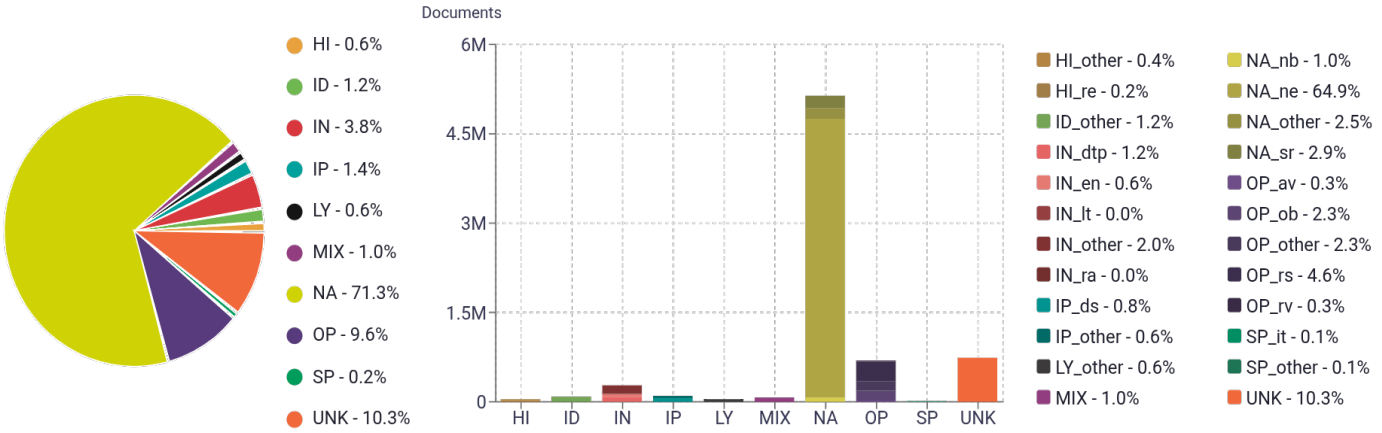
Top 10 domains

Domain	Docs	% of total
dailypakistan.c...	382K	5.30%
urdupoint.com	197K	2.73%
arynews.tv	162K	2.24%
jang.com.pk	143K	1.98%
nawaiwaqt.com.pk	140K	1.94%
express.pk	119K	1.66%
siasat.com	105K	1.46%
alarabiya.net	77K	1.07%
urduvoa.com	76K	1.05%
news18.com	67K	0.93%

Top 10 TLDs

Domain	Docs	% of total
com	3.6M	50.03%
com.pk	1M	14.25%
tv	680K	9.44%
pk	675K	9.36%
net	323K	4.48%
org	312K	4.33%
ir	62K	0.86%
in	54K	0.76%
site	32K	0.44%
info	31K	0.43%

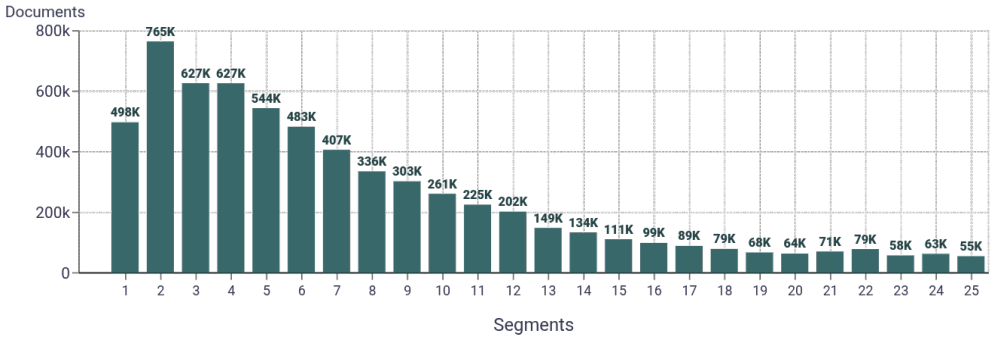
Register labels



MT:6.5% | 466K Documents

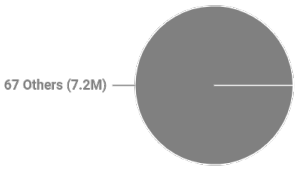
Documents size (in segments)

≤ 25 segments **88.73%** (6.4M documents)
> 25 segments **11.27%** (812K documents)



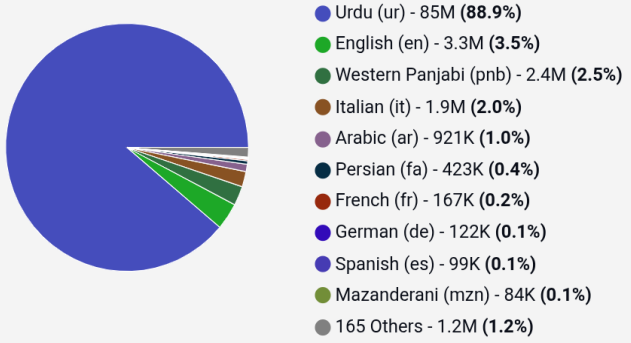
Document collections

CC = 94.33%
IA = 5.67%

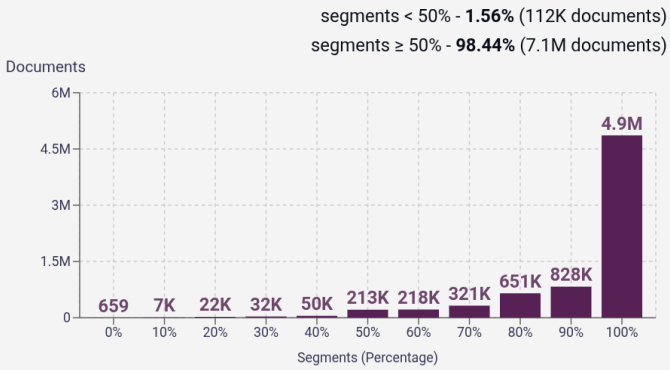


Language Distribution

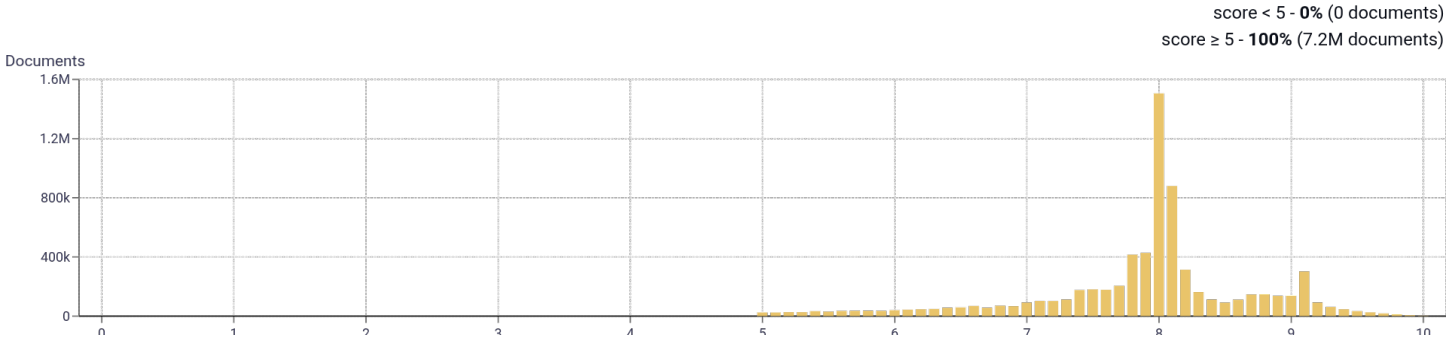
Number of segments in the Urdu (ur) corpus



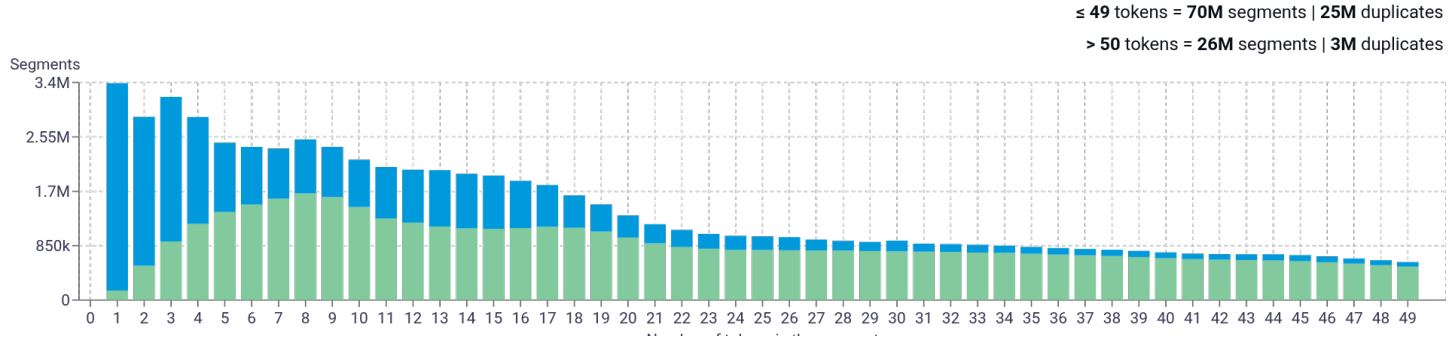
Percentage of segments in Urdu (ur) inside documents



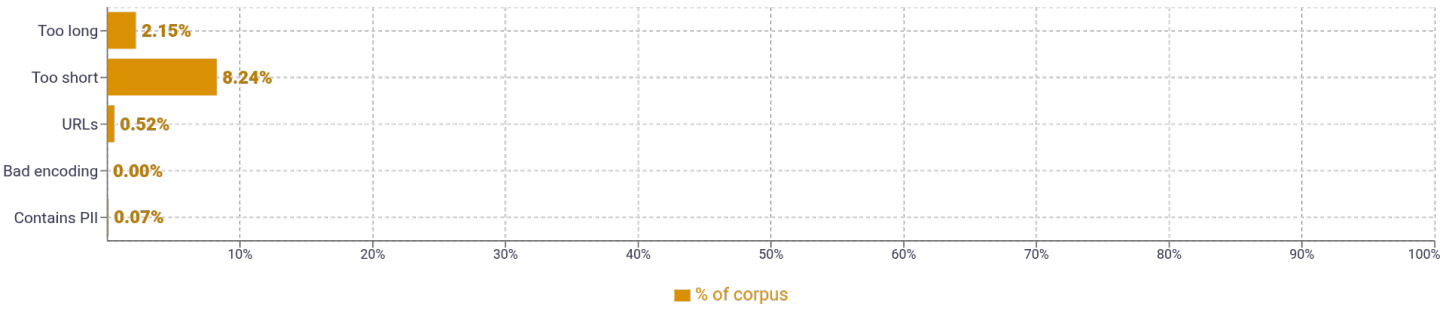
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	54,328,578 | نے · 44,229,674 | اس · 26,382,600 | بھی · 23,264,765 | کیا · 22,511,801 | کر
2	6,307,212 | نے کہا · 5,616,745 | انہوں نے · 3,200,076 | کیا گیا · 1,959,662 | کر دیا · 1,865,134 | عمران خان
3	1,892,440 | انہوں نے کہا · 1,397,672 | صلی اللہ علیہ · 1,205,766 | کرنے کے لیے · 1,038,518 | اللہ علیہ وسلم · 735,761 | اس کے بعد
4	1,019,214 | صلی اللہ علیہ وسلم · 367,668 | ان کا کہنا تھا · 356,181 | اللہ علیہ وسلم نے · 338,648 | اللہ صلی اللہ علیہ · 334,446 | رسول اللہ صلی اللہ
5	351,686 | صلی اللہ علیہ وسلم نے · 327,947 | رسول اللہ صلی اللہ علیہ · 289,537 | اللہ صلی اللہ علیہ وسلم · 205,938 | نبی کریم صلی اللہ علیہ · 204,186 | کریم صلی اللہ علیہ وسلم

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				