HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
urd_Arab.jsonl.tsv	9/22/2024	Urdu (ur)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
3,193,992	50,629,940	29,400,119 (58.07 %)	2.3B	9,958,862,188	16.4 GB

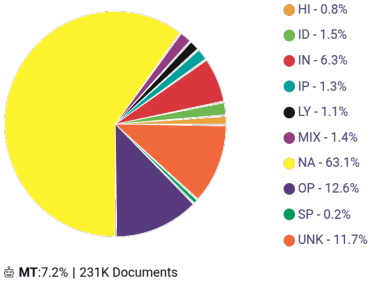
Top 10 domains

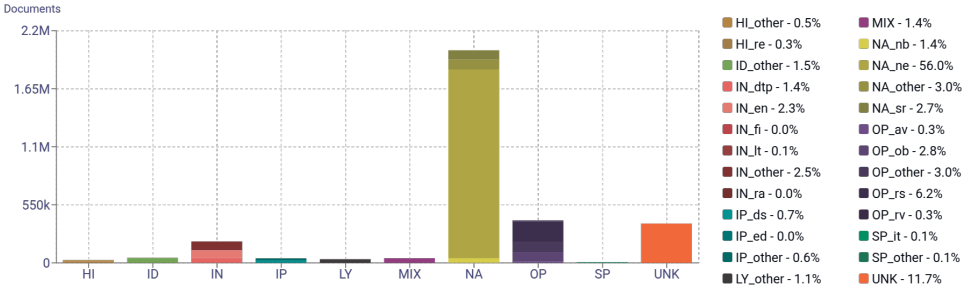
Domain	Docs	% of total
urduvoa.com	141K	4.40%
dailypakistan.c...	110K	3.44%
urdupoint.com	105K	3.28%
wikipedia.org	93K	2.91%
arynews.tv	85K	2.65%
siasat.com	72K	2.24%
nawaiwaqt.com.pk	58K	1.80%
geourdu.com	49K	1.54%
express.pk	39K	1.21%
news18.com	34K	1.08%

Top 10 TLDs

Domain	Docs	% of total
com	1.8M	55.04%
com.pk	348K	10.89%
org	225K	7.04%
tv	221K	6.92%
pk	217K	6.78%
net	127K	3.96%
info	31K	0.97%
in	25K	0.78%
xyz	23K	0.73%
ir	22K	0.70%

Register labels

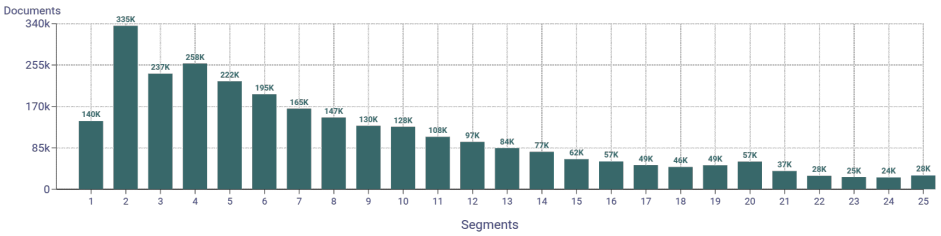




MT:7.2% | 231K Documents

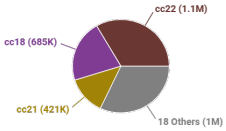
Documents size (in segments)

<= 25 segments **87.22%** (2.8M documents)
> 25 segments **12.78%** (408K documents)



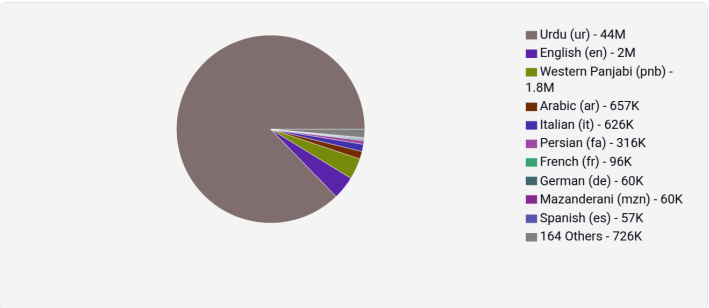
Documents by collection

CC = 78.09%
IA = 21.91%

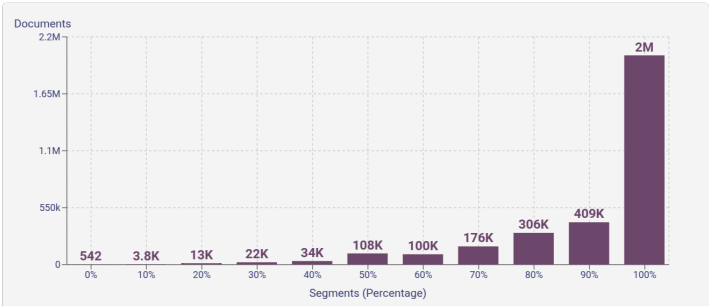


Language Distribution

Number of segments in the Urdu (ur) corpus

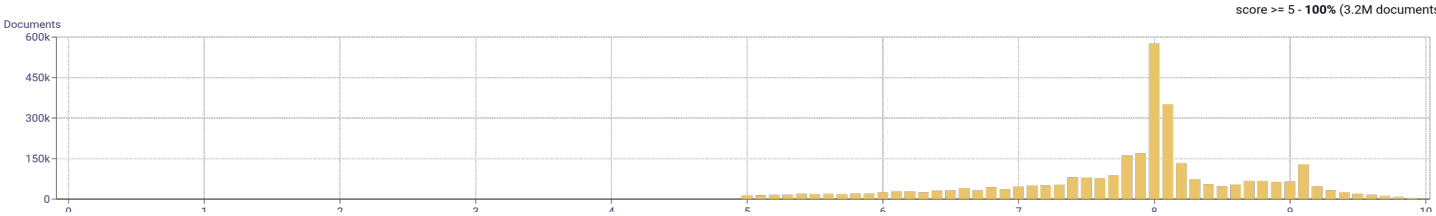


Percentage of segments in Urdu (ur) inside documents



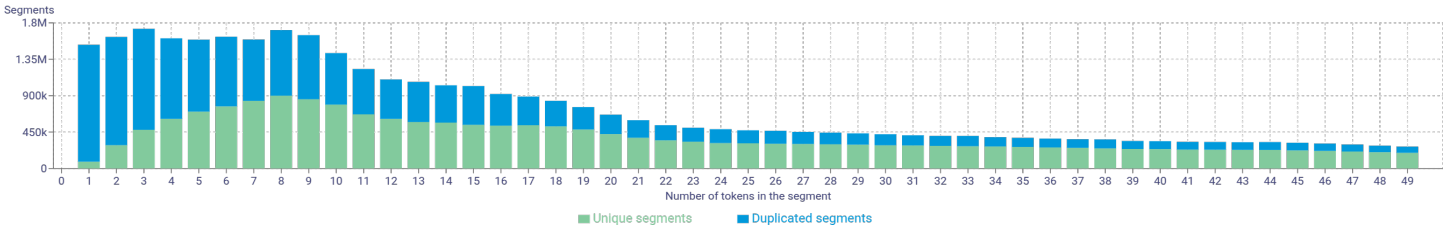
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (3.2M documents)

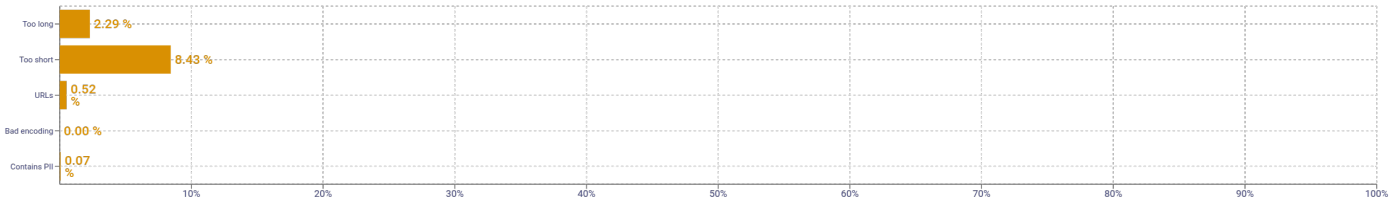


Segment length distribution by token

≤ 49 tokens = **20M** segments | **17M** duplicates
> 50 tokens = **14M** segments | **3.8M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	55733554 | میں · 40072337 | سے · 31036762 | کا · 30463430 | کو · 26414312 | نے
2	2633820 | نے کہا · 2442696 | انہوں نے · 1599731 | آپ کو · 1445997 | سب سے · 1445823 | اس کا
3	846072 | صلی اللہ علیہ · 729302 | انہوں نے کہا · 547056 | اللہ علیہ وسلم · 491060 | کا کہنا تھا · 479172 | کرنے کے لیے
4	531065 | صلی اللہ علیہ وسلم · 167532 | اللہ علیہ وسلم نے · 159620 | جس کی وجہ سے · 158120 | اللہ صلی اللہ علیہ · 157488 | اللہ علیہ وآلہ وسلم
5	163807 | صلی اللہ علیہ وسلم نے · 152812 | رسول اللہ صلی اللہ علیہ · 148815 | صلی اللہ علیہ وآلہ وسلم · 116547 | اللہ صلی اللہ علیہ وسلم · 84318 | آپ 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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av