HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

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

Top 10 domains

Domain	Docs	% of total
urduvoa.com	141K	4.40
dailypakistan.com.pk	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

Documents size (in segments)

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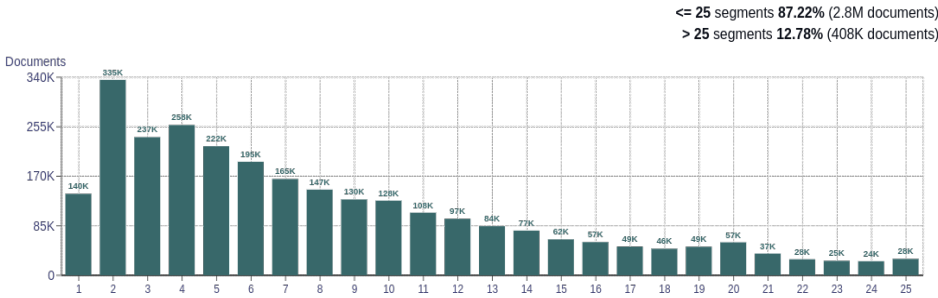
Documents by collection

cc22 (1.1M)
cc18 (685K)
cc21 (421K)
18 Others (1M)
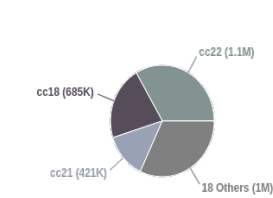
Language Distribution

Number of segments

- Urdu (ur) - 44M
- English (en) - 2M
- Western Panjabi (pnb) - 1.8M
- Arabic (ar) - 657K
- Italian (it) - 626K
- Persian (fa) - 316K
- French (fr) - 96K
- German (de) - 60K
- Mazanderani (mzn) - 60K
- Spanish (es) - 57K
- 164 Others - 726K
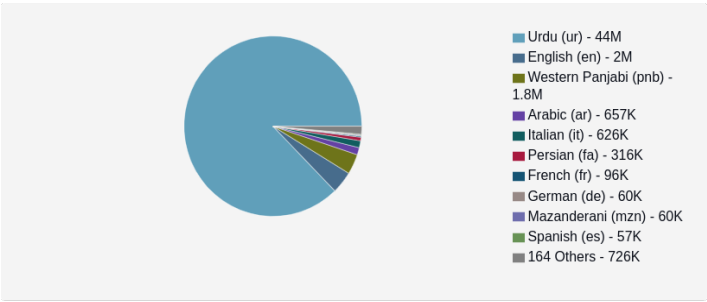
Percentage of segments in Urdu (ur) inside documents

Distribution of documents by document score

score <= 5 - **99.97%** (3.2M documents)
score > 5 - **0.03%** (960 documents)

Segment length distribution by token

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

Number of tokens in the segment

- Unique segments
- Duplicated segments

Segment noise distribution

- Too long — 0.00 %
- Too short — 3.95 %
- URLs — 0.39 %
- Bad encoding — 0.00 %
- Contains PII — 0.07 %
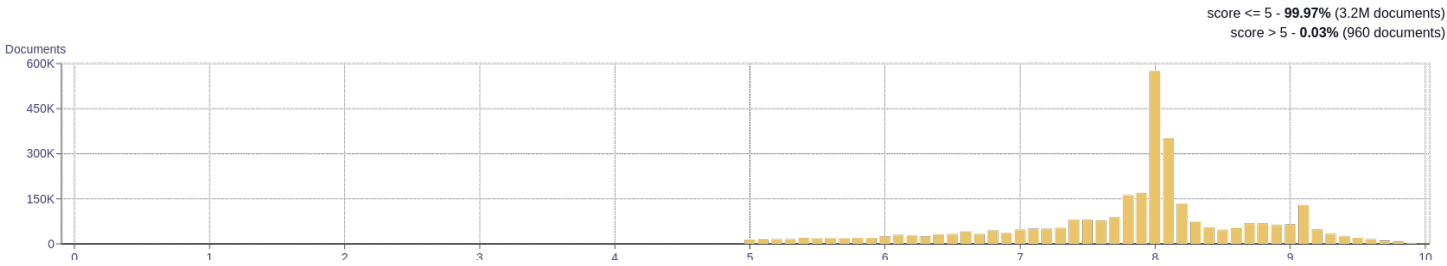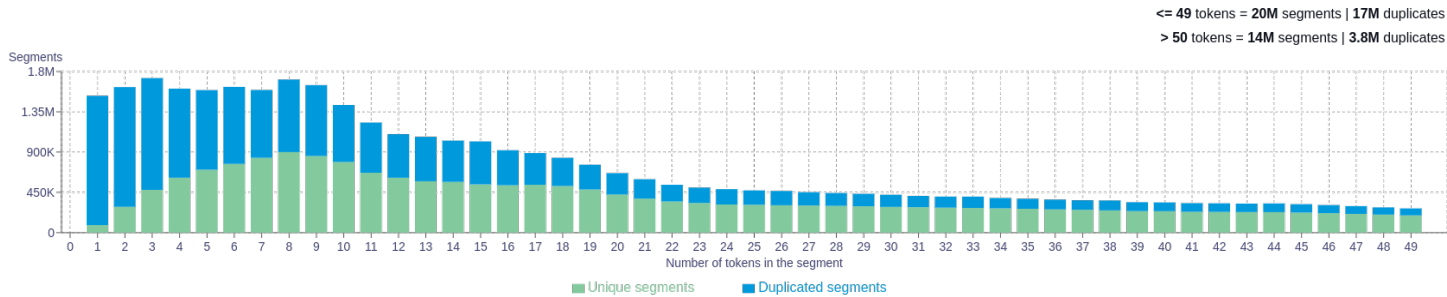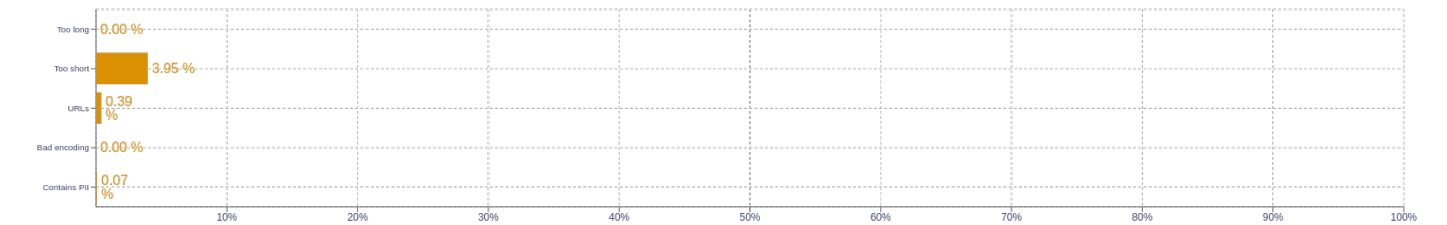
Frequent n-grams

Size	n-grams
1	میں | 55733554 سے | 40072337 کا | 31036762 کو | 30463430 نے | 26414312
2	اس کا | 1445823 سے سے | 1445997 آپ کو | 1599731 انہوں نے | 2442696 نے کہا | 2633820
3	کرتے کہ لیے | 479172 کا کہنا تھا | 491060 اللہ علیہ وسلم | 547056 انہوں نے کہا | 729302 صلی اللہ علیہ | 846072
4	اللہ علیہ وآلہ وسلم | 157488 اللہ صلی اللہ علیہ | 158120 جس کی وجہ سے | 159620 اللہ علیہ وسلم نے | 167532 صلی اللہ علیہ وسلم | 531065
5	آپ صلی اللہ علیہ وسلم | 84318 اللہ صلی اللہ علیہ وسلم | 116547 صلی اللہ علیہ وآلہ وسلم | 148815 رسول اللہ صلی اللہ علیہ | 152812 صلی اللہ علیہ وسلم نے | 163807

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