

General overview

Corpus	Analytics date	Language
pbt_Arab.jsonl.tsv	9/20/2024	Southern Pashto (pbt)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
466,472	8,454,662	5,506,825 (65.13 %)	306M	2.13 GB	1,295,601,474

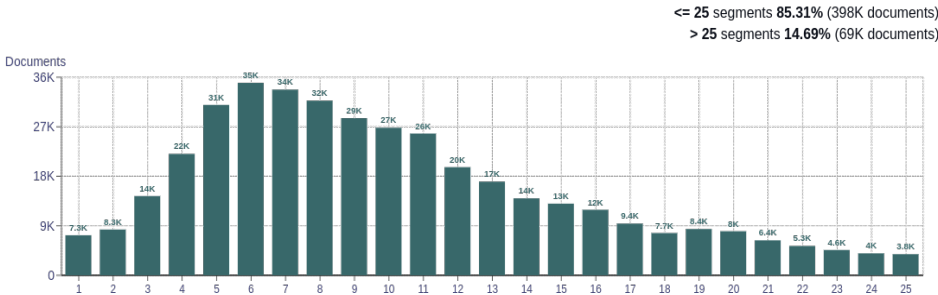
Top 10 domains

Domain	Docs	% of total
pashতোবা.com	38K	8.10
mashaalradio.com	32K	6.81
bakhtarnews.af	28K	6.01
tolonews.com	24K	5.23
nunn.asia	15K	3.21
tolaafghan.com	13K	2.89
wikipedia.org	12K	2.63
larawbar.net	12K	2.48
taand.com	11K	2.44
dw.com	10K	2.25

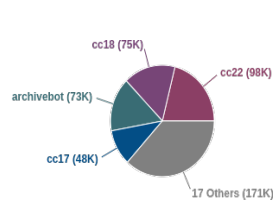
Top 10 TLDs

Domain	Docs	% of total
com	288K	61.68
af	49K	10.45
gov.af	25K	5.33
net	24K	5.11
org	23K	4.93
asia	15K	3.23
cn	6.8K	1.46
website	4.6K	0.98
info	3.5K	0.74
ir	2.7K	0.57

Documents size (in segments)

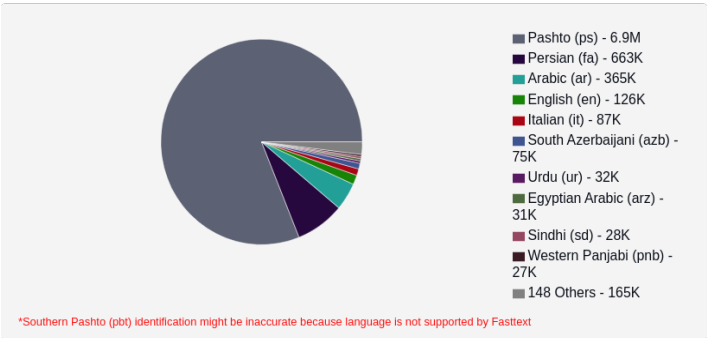


Documents by collection

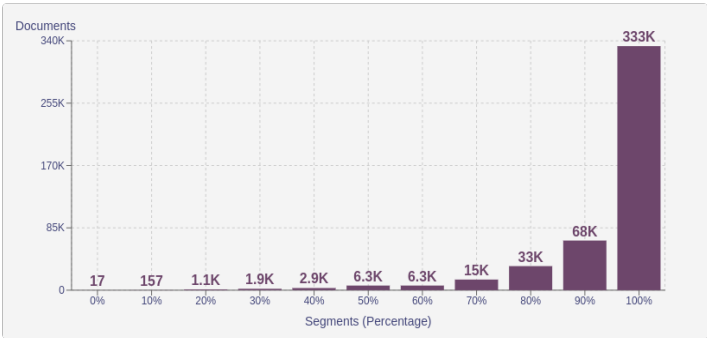


Language Distribution

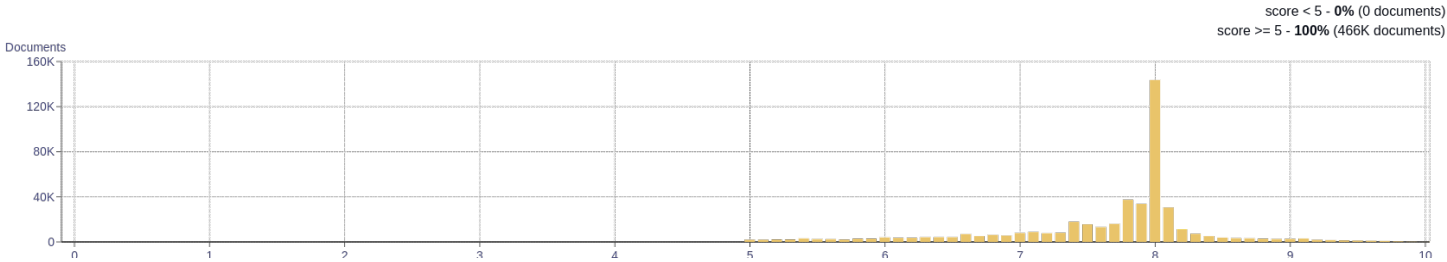
Number of segments



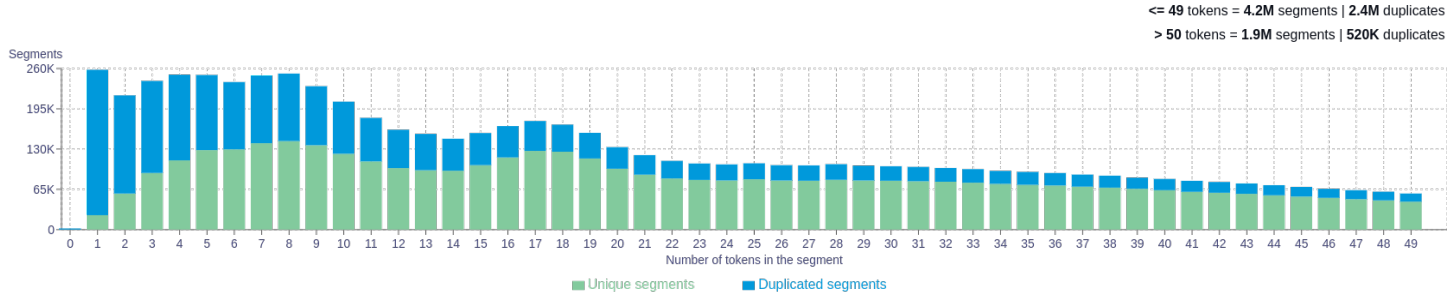
Percentage of segments in Southern Pashto (pbt) inside documents



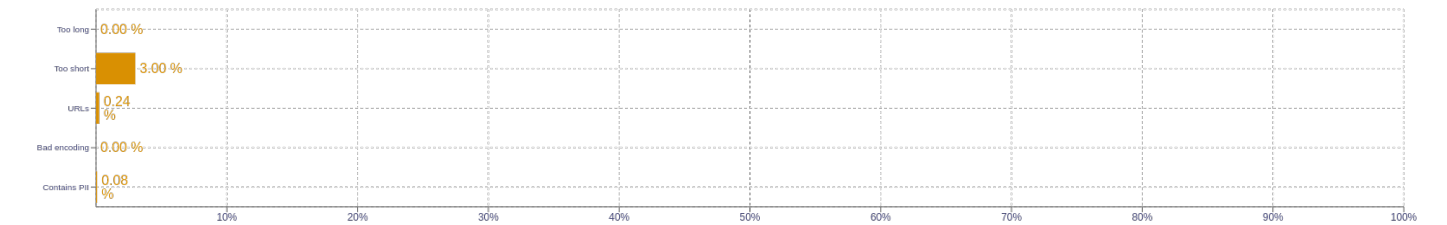
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	بي 1341005 دي 1624373 دي 1706720 كي 5332453 چي 5754718
2	حال كي 110752 كي چي 116629 برخه كي 130557 دي چي 188791 افغانستان كي 193226
3	چي د دي 31550 كي د افغانستان 36106 حال كي چي 39227 چي د افغانستان 51729 چي په دي 56303
4	صلی الله عليه وسلم 13877 افغانستان د اسلامي جمهوريت 14219 صلی الله عليه وسلم 16442 صلی الله عليه وسلم 25413 چي په افغانستان كي 28146
5	کې د پښتنو او بلوڅو 4178 جمهور رئيس محمد اشرف غني 4799 رسول الله صلی الله عليه 7995 رسول الله صلی الله عليه 8720 رسول الله صلی الله عليه 12440

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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>