HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
ps_1.jsonl.tsv	3/17/2024	Pashto (ps)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
88,212	10,984,513	2,078,407 (18.92 %)	131M	900.22 MB	

Top 10 domains

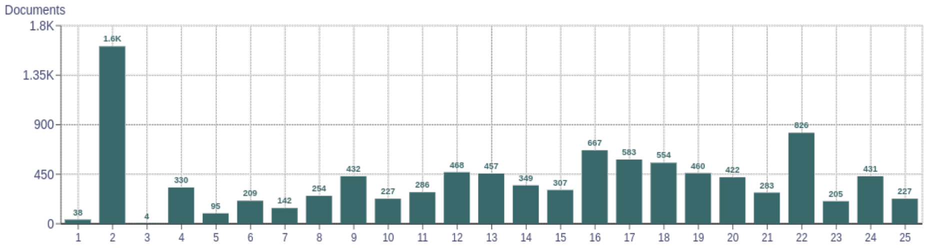
Domain	Docs	% of total
taand.com	3.8K	4.27
spogmairadio.af	3.5K	4.00
larawbar.net	2.7K	3.04
qamosona.com	2.5K	2.87
pashtovoa.com	2.2K	2.52
khabarial.com	2.2K	2.46
kandahar-tv.com	2.1K	2.42
cri.cn	1.9K	2.16
nunn.asia	1.8K	2.04
wikipedia.org	1.7K	1.92

Top 10 TLDs

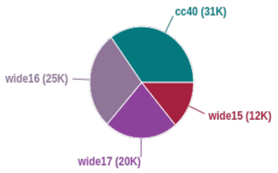
Domain	Docs	% of total
com	55K	62.31
af	8K	9.03
net	6.1K	6.97
org	5K	5.70
gov.af	2.6K	2.99
cn	1.9K	2.16
asia	1.9K	2.11
info	1.5K	1.70
org.af	570	0.65
tv	515	0.58

Documents size (in segments)

<= 25 segments **11.19%** (9.9K documents)
> 25 segments **88.81%** (78K documents)

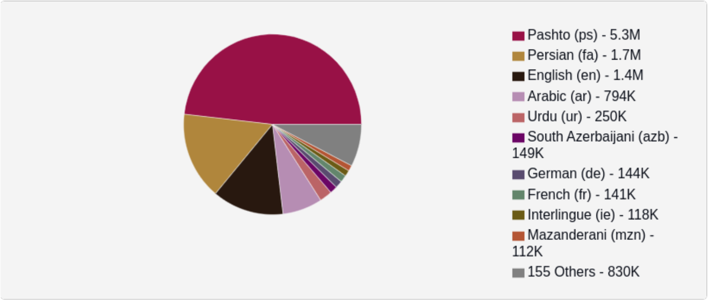


Documents by collection

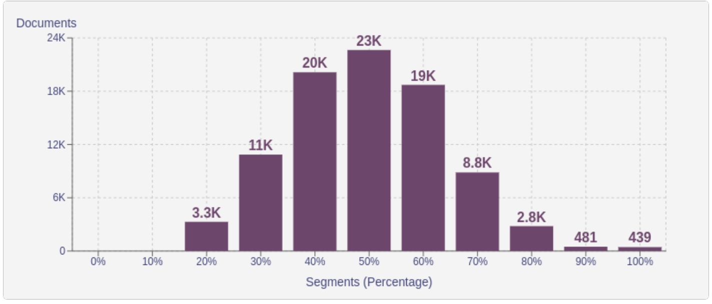


Language Distribution

Number of segments

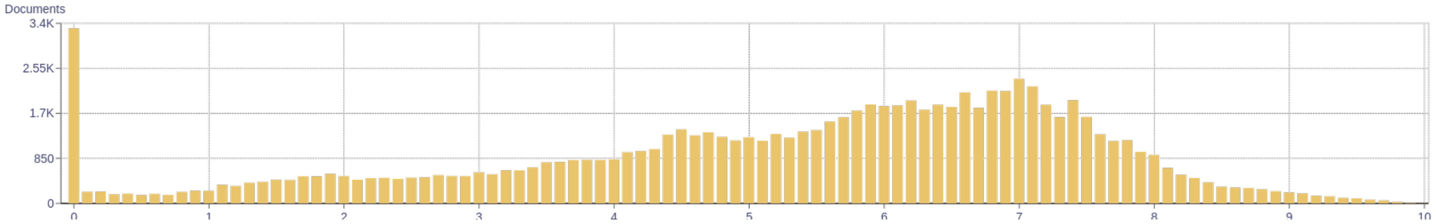


Percentage of segments in Pashto (ps) inside documents



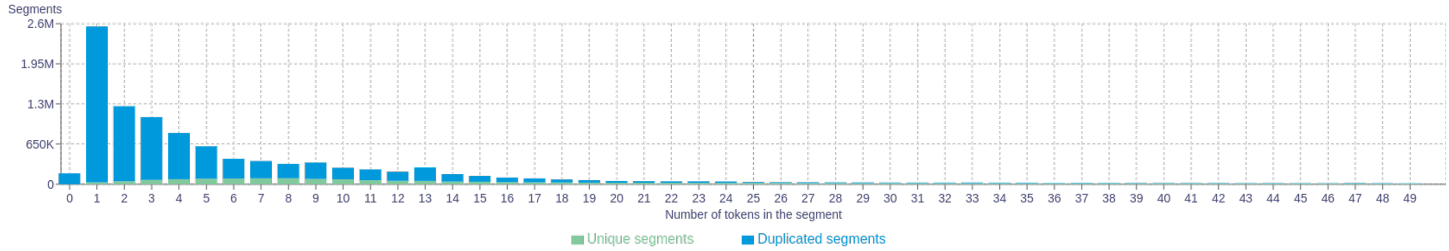
Distribution of documents by document score

score <= 5 - **37.17%** (33K documents)
score > 5 - **62.83%** (55K documents)

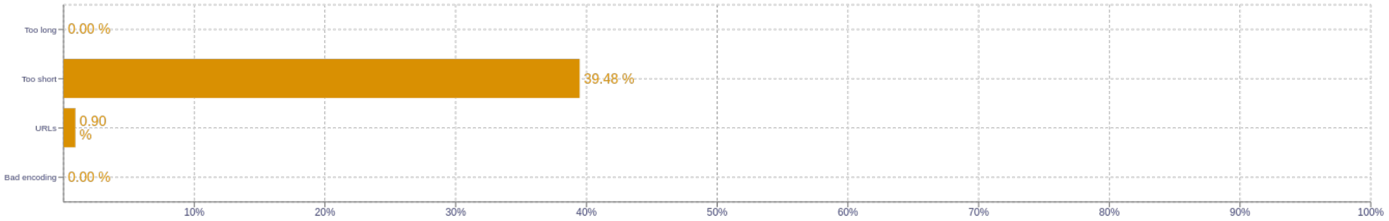


Segment length distribution by token

<= 49 tokens = **1.7M** segments | **8.8M** duplicates
> 50 tokens = **509K** segments | **111K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	1652289 | کي 1388817 | چې 501104 | دي 464528 | افغانستان 428709 | دې
2	71463 | افغانستان کي 51888 | مهم خبرونه 46617 | اونۍ مهم days ago | 43757 hours ago | 42234
3	46614 | اونۍ مهم خبرونه all rights reserved | 21826 20739 | داسي هم شته 13008 | صلی الله علیه 11728 | ساینس او ټکنالوژي
4	11980 | صلی الله علیه وسلم 9078 | دنور لارویان ږغیزه خبرونه 6580 | چې په افغانستان کي from twitter for iphone | 6146 opens in new window | 5917
5	5450 | رسول الله صلی الله علیه a password will be e | 4498 4343 | مونږ سره په تماس کي your email address will not | 3816 email address will not be | 3816

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