HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-docslite.fa.tsv	7/2/2024	Persian (fa)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
30,897,583	4,865,748,733			403.06 GB	

Top 10 domains

Domain	Docs	% of total
blogfa.com	557K	1.80
portalekhabar.ir	314K	1.02
mihanblog.com	243K	0.79
ex1rs.com	227K	0.74
khabarfarsi.com	198K	0.64
blog.ir	170K	0.55
tnews.ir	144K	0.47
parscenter.com	133K	0.43
akairan.com	122K	0.40
afarineshdaily.ir	122K	0.39

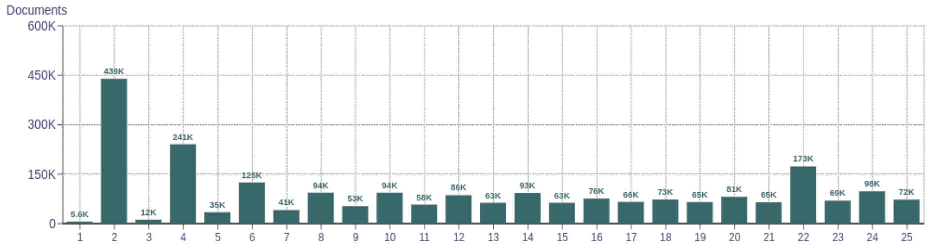
Top 10 TLDs

Domain	Docs	% of total
ir	14M	46.39
com	13M	40.53
net	725K	2.35
org	640K	2.07
ac.ir	247K	0.80
in	216K	0.70
gdn	212K	0.69
xyz	182K	0.59
pl	150K	0.49
nl	141K	0.45

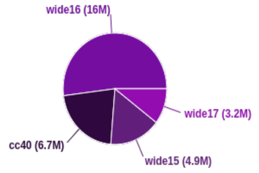
Documents size (in segments)

<= 25 segments **7.58%** (2.3M documents)

> 25 segments **92.42%** (29M documents)

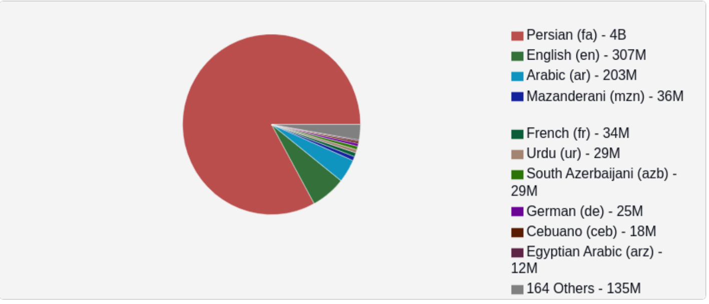


Documents by collection

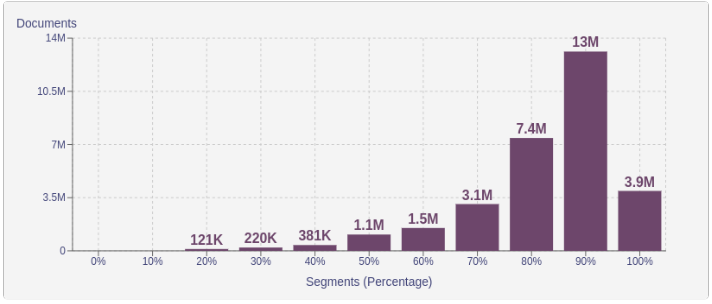


Language Distribution

Number of segments



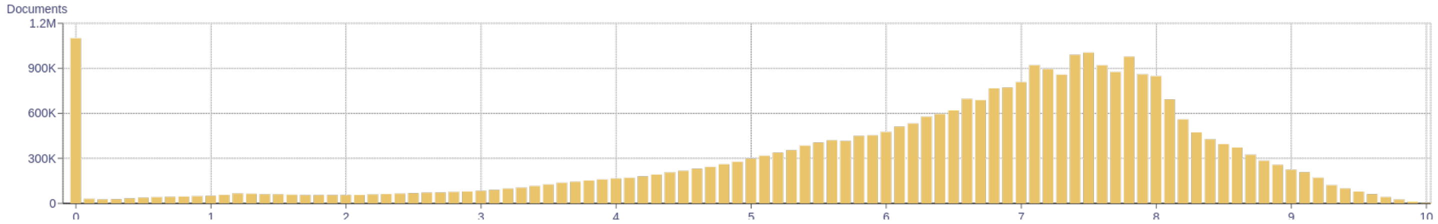
Percentage of segments in Persian (fa) inside documents



Distribution of documents by document score

score <= 5 - **19.51%** (6M documents)

score > 5 - **80.49%** (25M documents)



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