HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-v2-pes_Arab.tsv	9/26/2024	Persian (pes)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
90,498,985	3,963,188,468			740.98 GB	451,189,565,359

Top 10 domains

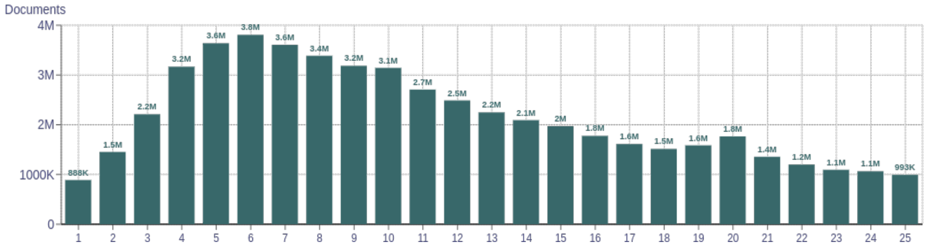
Domain	Docs	% of total
blogfa.com	8.5M	9.40
netct.ir	2.8M	3.13
netgarmi.in	2.7M	3.03
netgarmi.ir	2.5M	2.80
patoghy.ir	1.8M	2.03
persianblog.ir	1.2M	1.35
mihanblog.com	898K	0.99
blogsky.com	539K	0.60
blog.ir	512K	0.57
akairan.com	507K	0.56

Top 10 TLDs

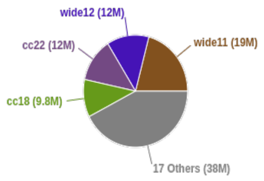
Domain	Docs	% of total
ir	41M	45.59
com	38M	42.20
in	3M	3.27
net	2.3M	2.51
org	2.1M	2.37
ac.ir	504K	0.56
info	231K	0.26
xyz	176K	0.19
co	146K	0.16
nl	138K	0.15

Documents size (in segments)

<= 25 segments **59.6%** (54M documents)
> 25 segments **40.4%** (37M documents)

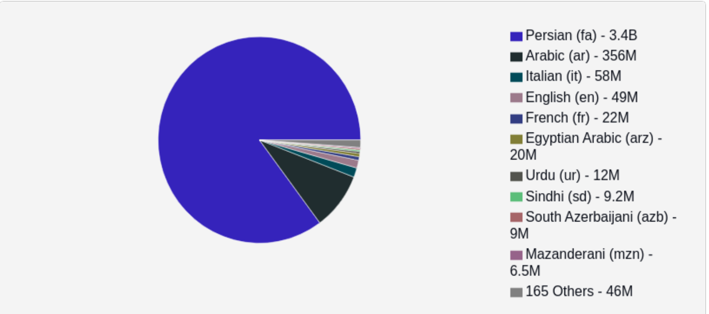


Documents by collection



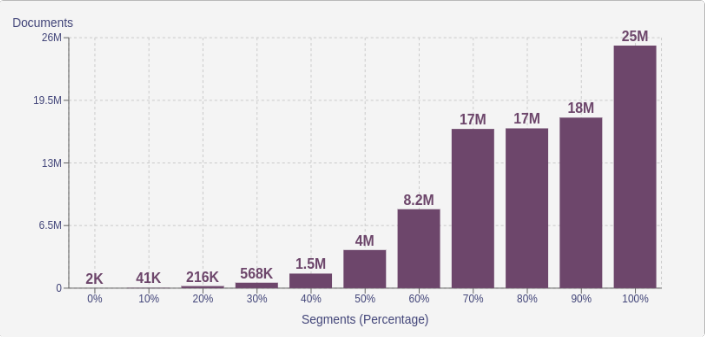
Language Distribution

Number of segments



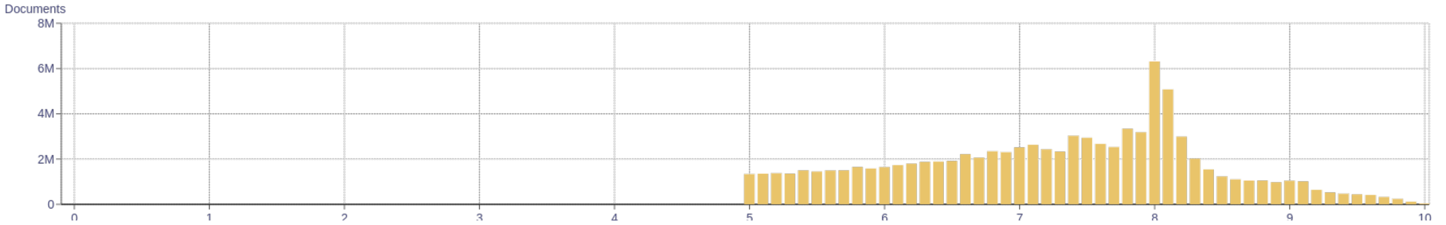
*Persian (pes) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Persian (pes) inside documents

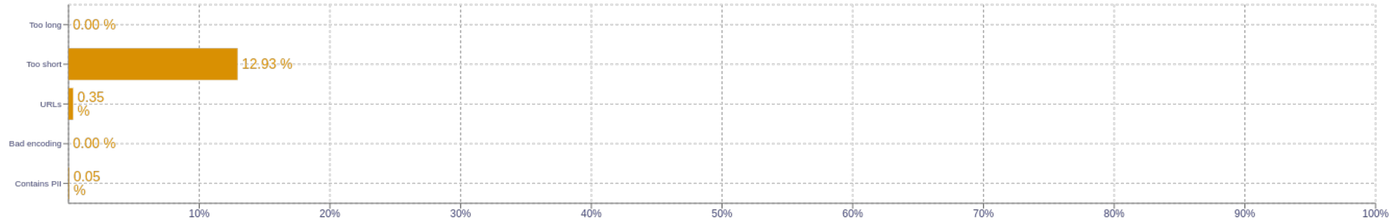


Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (90M documents)



Segment noise distribution



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