HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-pes_Arab	10/28/2025	Iranian Persian

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
126,990,828	3,863,122,318	1,584,813,539 (41.02 %)	58.98%	112B	501,865,492,688	824.42 GB

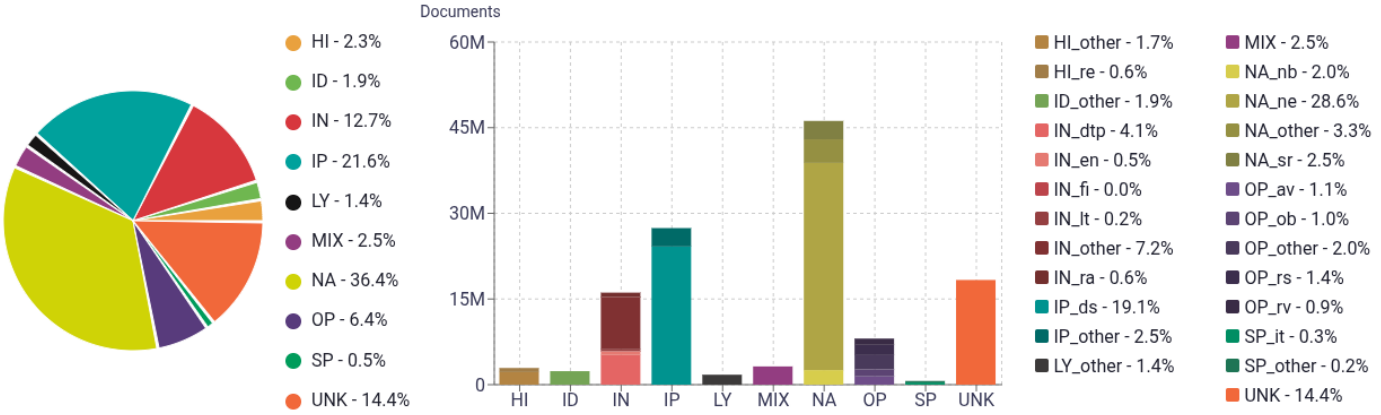
Top 10 domains

Domain	Docs	% of total
blogfa.com	5.1M	4.03%
netct.ir	2.8M	2.17%
netgarmi.in	2.5M	1.94%
netgarmi.ir	2.4M	1.91%
patoghy.ir	1.6M	1.30%
mihanblog.com	937K	0.74%
persianblog.ir	899K	0.71%
blog.ir	490K	0.39%
jafo.ir	452K	0.36%
khabarfarsi.com	450K	0.35%

Top 10 TLDs

Domain	Docs	% of total
ir	58M	45.56%
com	48M	38.16%
net	3.1M	2.41%
in	2.8M	2.17%
org	2.7M	2.10%
pl	1.6M	1.29%
nl	1.4M	1.08%
de	1M	0.80%
ac.ir	979K	0.77%
news	615K	0.48%

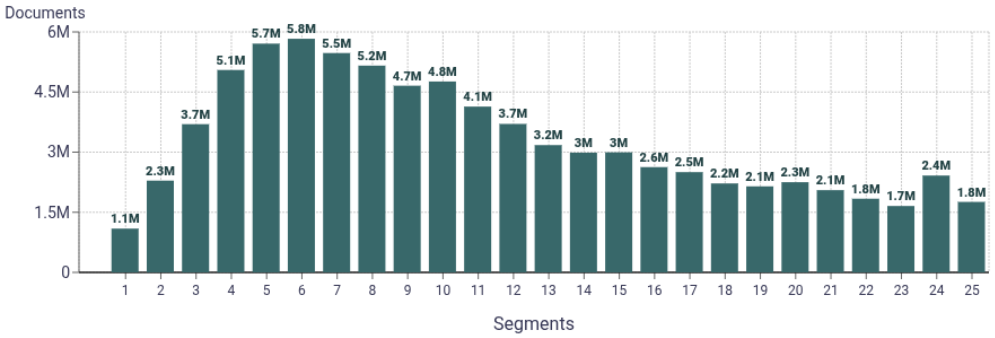
Register labels



MT:5.1% | 6.5M Documents

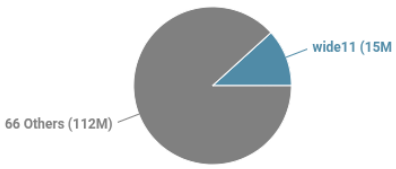
Documents size (in segments)

≤ 25 segments **64.72%** (82M documents)
> 25 segments **35.28%** (45M documents)



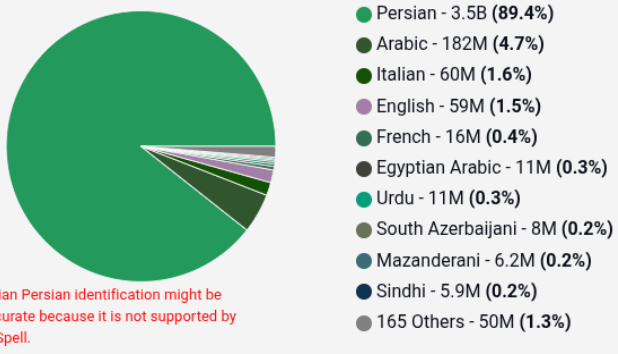
Document collections

CC = 67.76%
IA = 32.24%



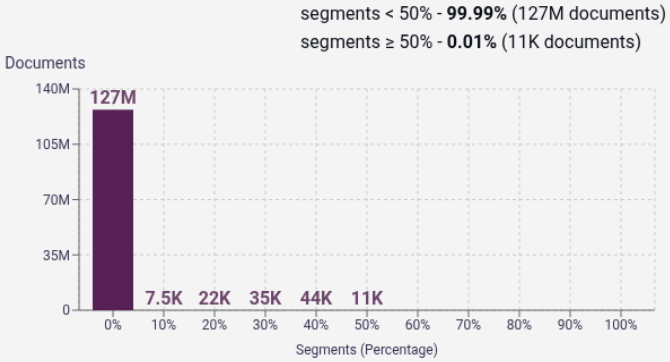
Language Distribution

Number of segments in the Iranian Persian corpus

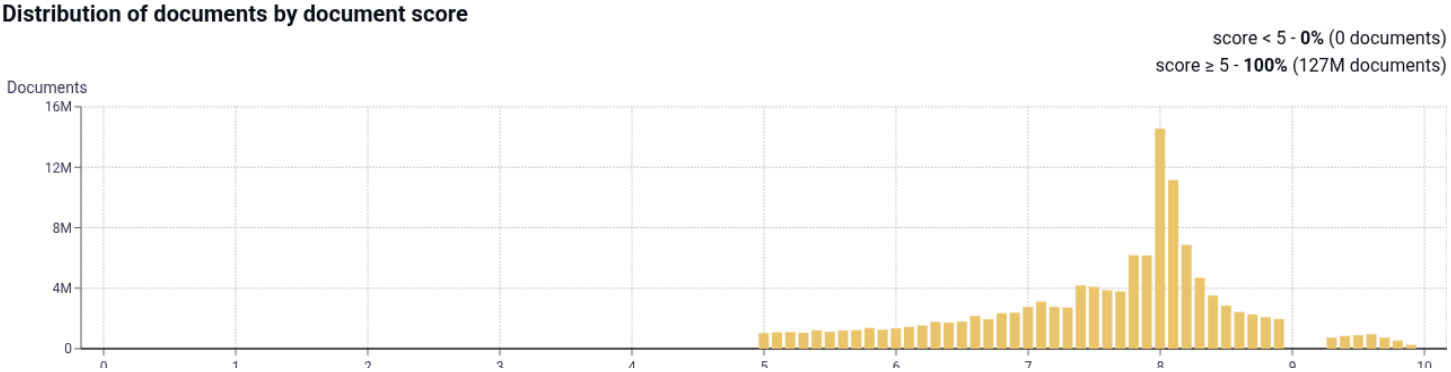


*Iranian Persian identification might be inaccurate because it is not supported by FastSpell.

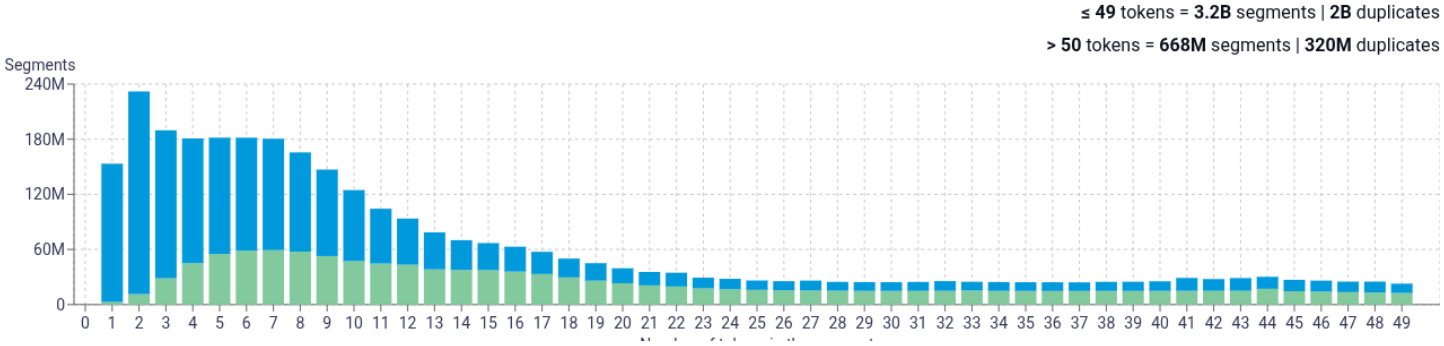
Percentage of segments in Iranian Persian inside documents



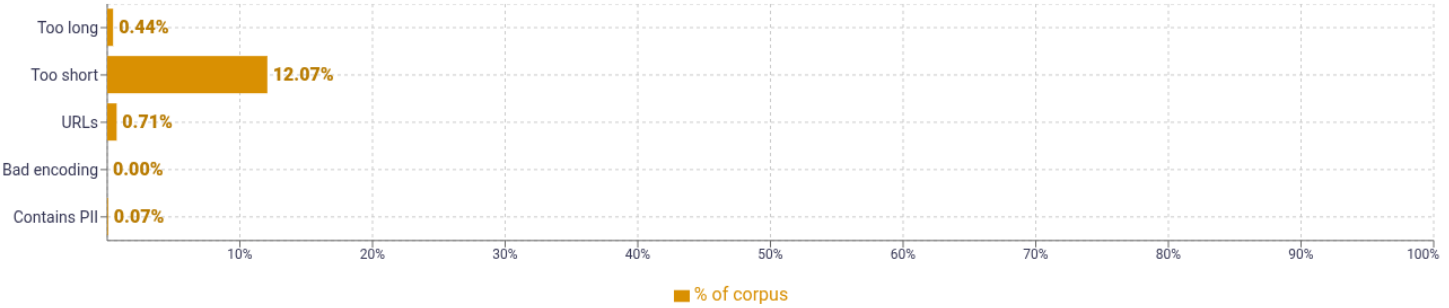
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	این | 1,218,570,041 · می | 802,559,990 · های | 729,177,550 · برای | 708,592,804 · سنگ | 316,229,034
2	سنگ شکن | 162,103,476 · ادامه مطلب | 91,483,136 · می کند | 63,786,576 · دانلود آهنگ | 45,102,412 · نرم افزار | 41,943,152
3	شن و ماسه | 20,338,185 · دانلود آهنگ جدید | 17,647,533 · سنگ شکن فکی | 17,598,280 · اس ام اس | 13,752,154 · سنگ شکن سنگ | 13,412,424
4	دستگاه های سنگ شکن | 8,396,332 · http www topseda ir | 6,326,602 · rel nofollow href http | 6,326,600 · nofollow href http www | 6,326,600 · href http www topseda | 6,326,600
5	rel nofollow href http www | 6,326,600 · nofollow href http www topseda | 6,326,600 · href http www topseda ir | 6,326,600 · a rel nofollow href http | 6,326,600 · علمی و آموزشی تاریخ انتشار | 5,409,316

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

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	**HI**	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	**Informational persuasion**	**IP**	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	**Opinion**	**OP**
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	**Informational description**	**IN**	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				