

General overview

Corpus	Date	Language
hplt-v3-prs_Arab	9/24/2025	Dari (prs)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
2,457,581	47,480,759	34,931,281 (73.57 %)	1.4B	6,346,471,447	10.24 GB

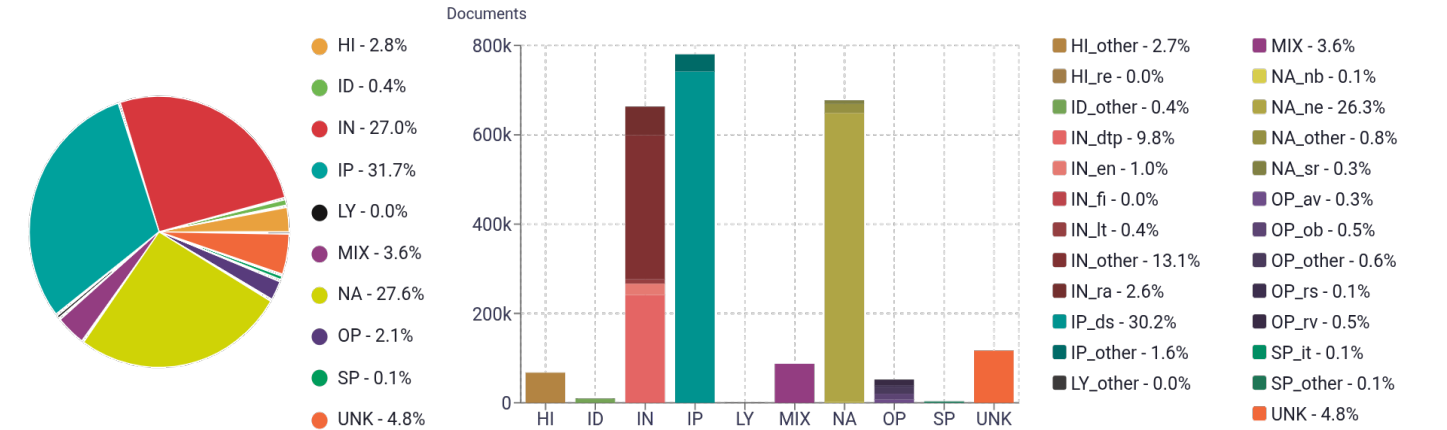
Top 10 domains

Domain	Docs	% of total
sputniknews.com	39K	1.58%
darivoa.com	39K	1.57%
darinews.com	34K	1.37%
azadiradio.com	25K	1.02%
blogfa.com	23K	0.95%
avapress.com	19K	0.78%
wikipedia.org	15K	0.61%
ariananews.co	15K	0.59%
mandegardaily.com	13K	0.52%
ariananews.af	13K	0.52%

Top 10 TLDs

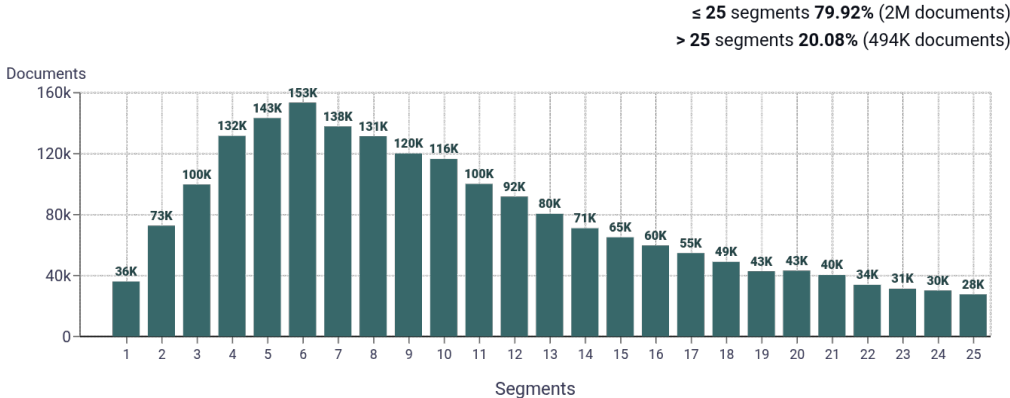
Domain	Docs	% of total
com	1.3M	51.85%
ir	731K	29.73%
af	85K	3.46%
org	71K	2.87%
net	66K	2.68%
gov.af	34K	1.40%
ac.ir	33K	1.33%
co	28K	1.15%
news	11K	0.47%
com.af	11K	0.43%

Register labels

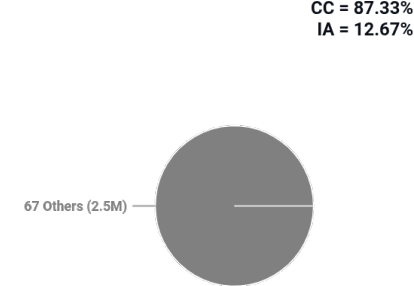


MT:1.5% | 37K Documents

Documents size (in segments) ⓘ

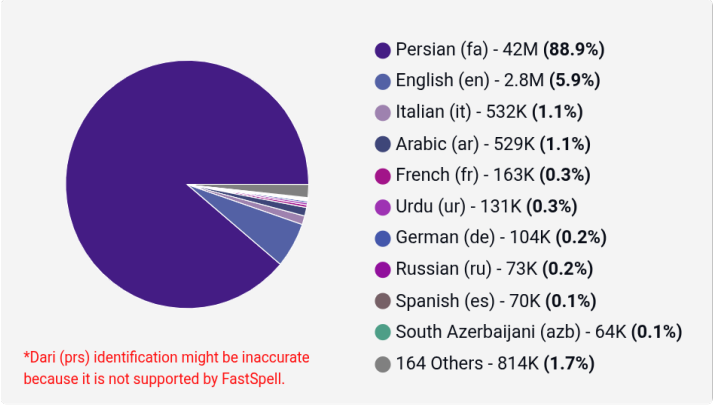


Document collections

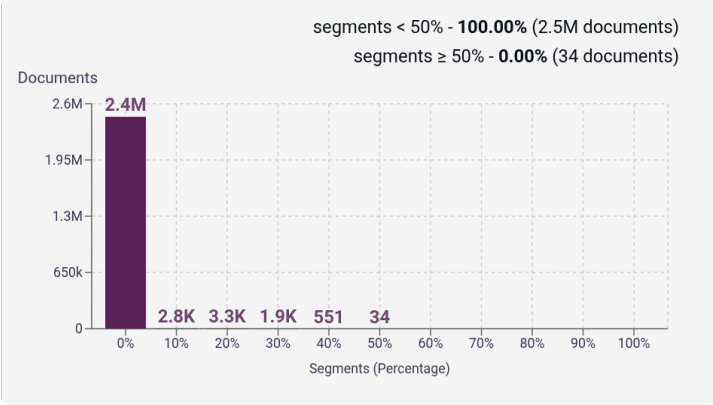


Language Distribution

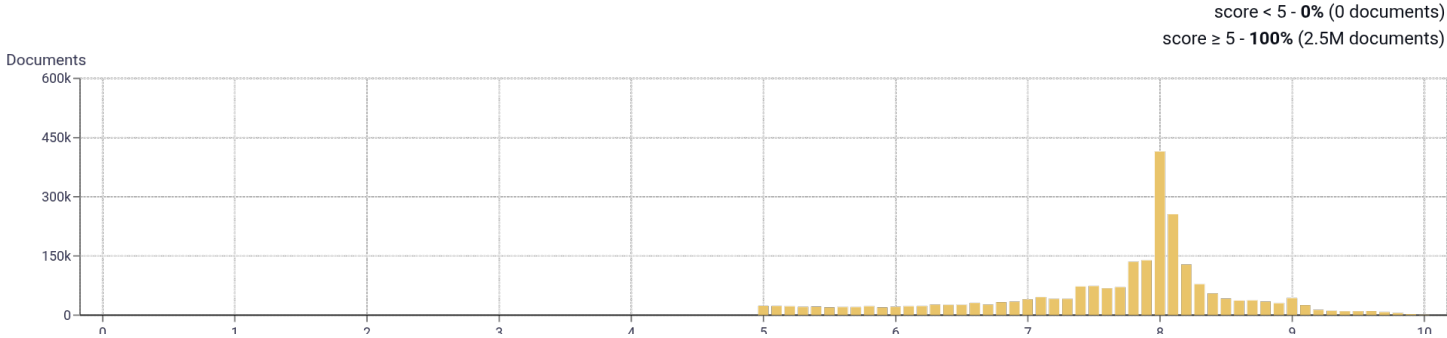
Number of segments in the Dari (prs) corpus



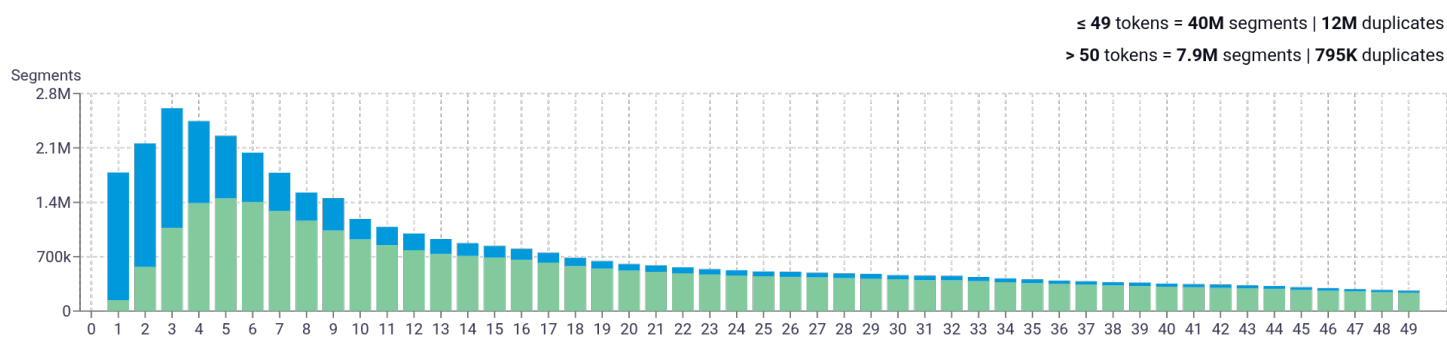
Percentage of segments in Dari (prs) inside documents



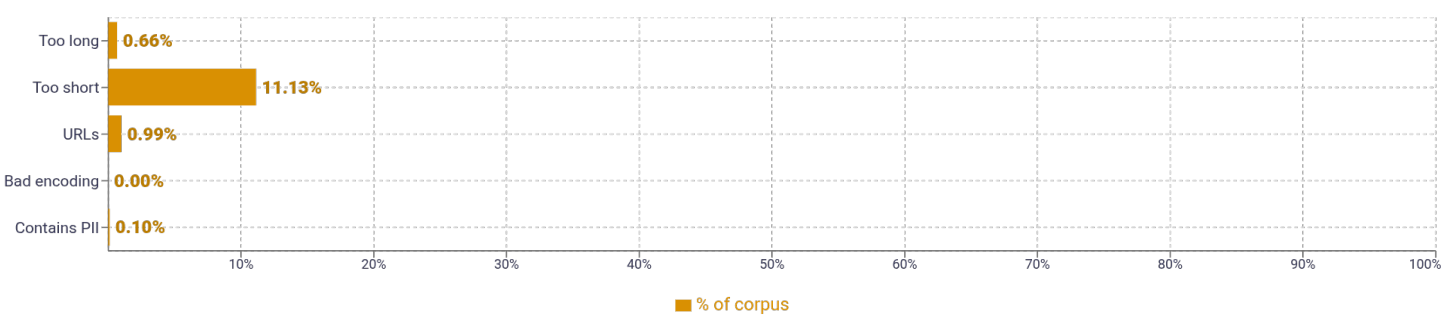
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	8,014,613   می   4,008,353   میشود   2,765,329   افغانستان   2,398,663   فرار   2,332,398   کند	
2	717,558   می کند   412,233   نرم افزار   306,775   نشان میدهد   302,933   ایالات متحده   294,332   هوش مصنوعی	
3	148,477   تجزیه و تحلیل   101,140   ۰۰۰   98,713   منحصر به فرد   96,626   سازمان ملل متحد   94,417   نقد و بررسیها	
4	115,785   دیدگاهی برای این محصول   98,980   ۰۰۰۰   44,901   صد عفونی کننده دست   24,352   تامین کننده تماس بگیردوانس   23,348   نصب و راه اندازی	
5	115,553   دیدگاهی برای این محصول نوشته   96,887   ۰۰۰۰۰۰   20,801   مشخصات مقاله ترجمه عنوان مقاله   18,912   صفحه هزینه دانلود مقاله انگلیسی   18,290   هزینه دانلود مقاله انگلیسی رایگان	

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 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>, <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				