HPLT Analytics report

HPLTAnalytics

General overview

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

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

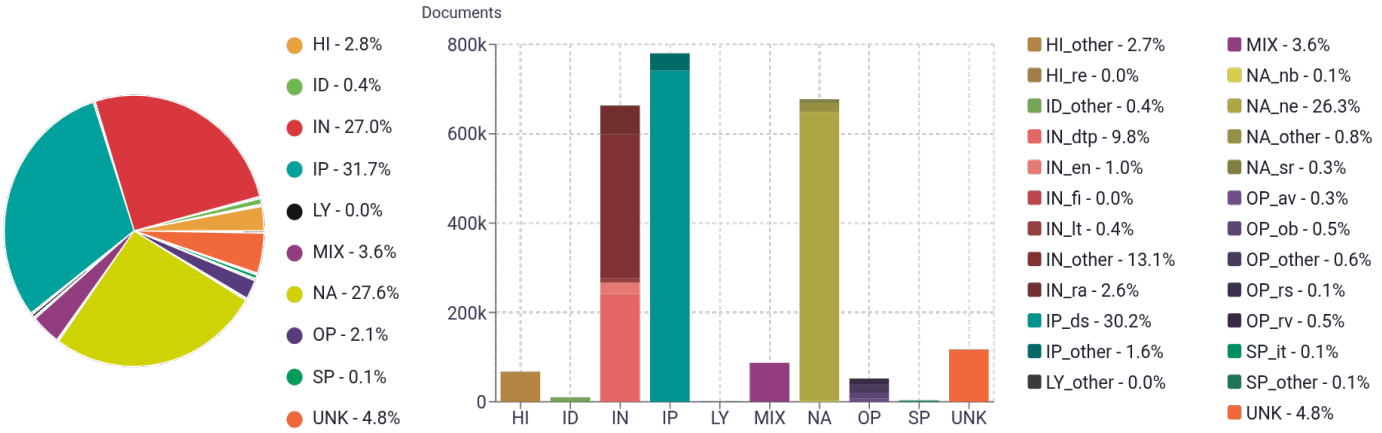
Top 10 domains

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%

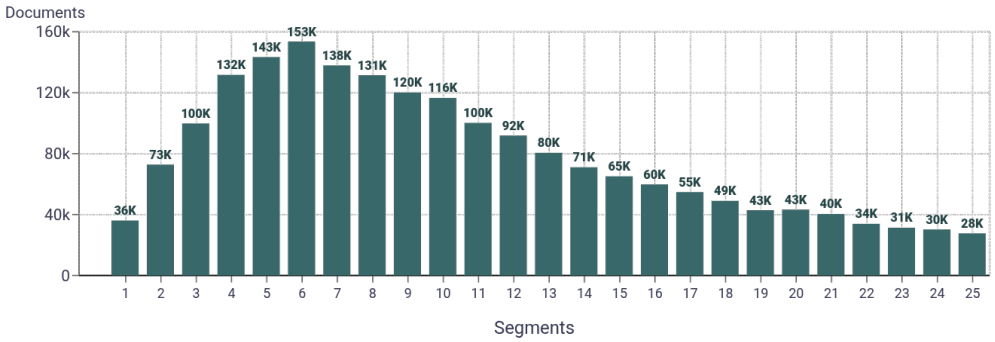
Register labels



MT:1.5% | 37K Documents

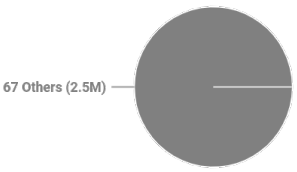
Documents size (in segments)

≤ 25 segments **79.92%** (2M documents)
> 25 segments **20.08%** (494K documents)



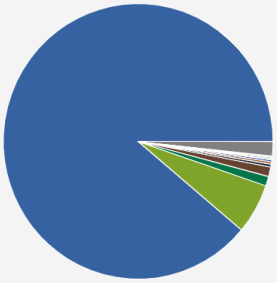
Document collections

CC = 87.33%
IA = 12.67%



Language Distribution

Number of segments in the Dari corpus

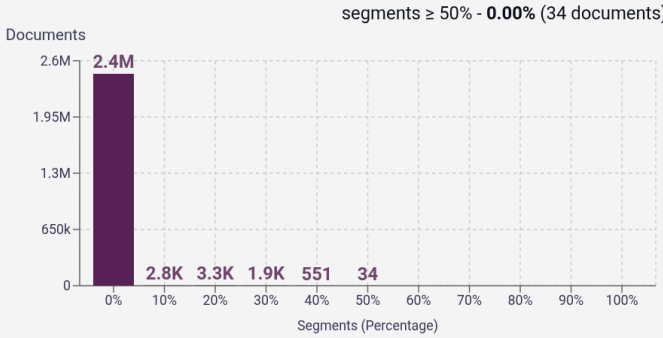


- Persian - 42M **(88.9%)**
- English - 2.8M **(5.9%)**
- Italian - 532K **(1.1%)**
- Arabic - 529K **(1.1%)**
- French - 163K **(0.3%)**
- Urdu - 131K **(0.3%)**
- German - 104K **(0.2%)**
- Russian - 73K **(0.2%)**
- Spanish - 70K **(0.1%)**
- South Azerbaijani - 64K **(0.1%)**
- 164 Others - 814K **(1.7%)**

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

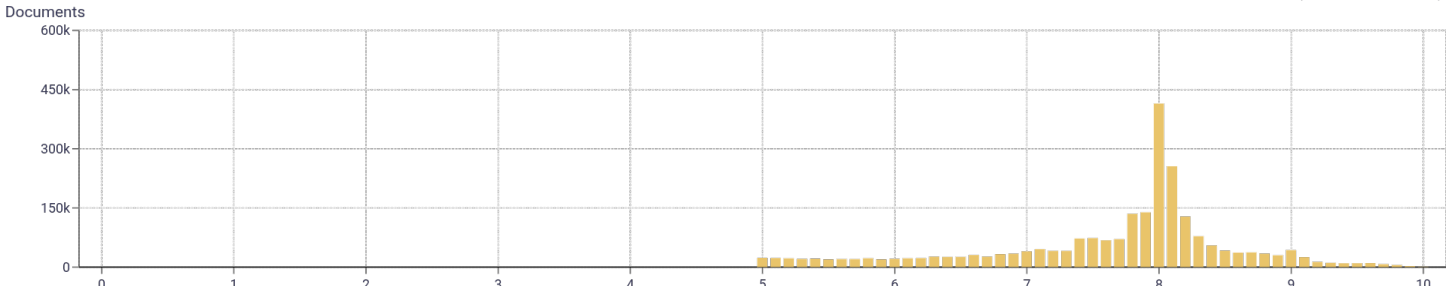
Percentage of segments in Dari inside documents

segments < 50% - **100.00%** (2.5M documents)
segments ≥ 50% - **0.00%** (34 documents)



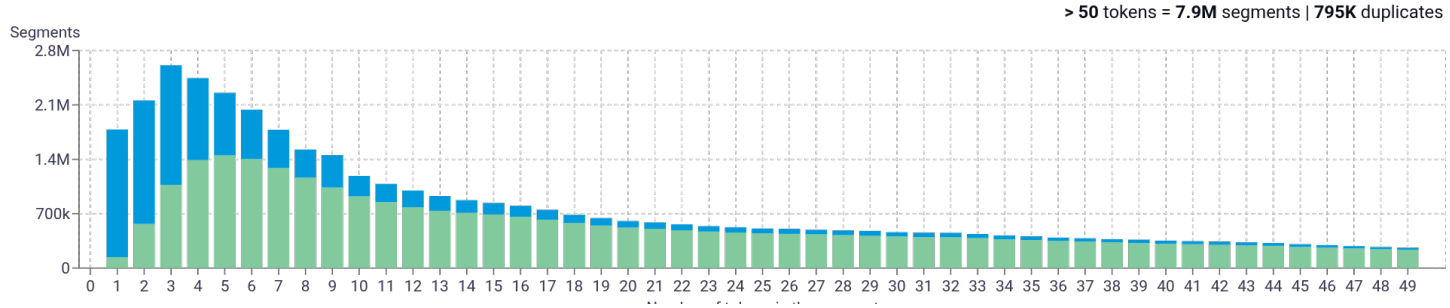
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (2.5M documents)

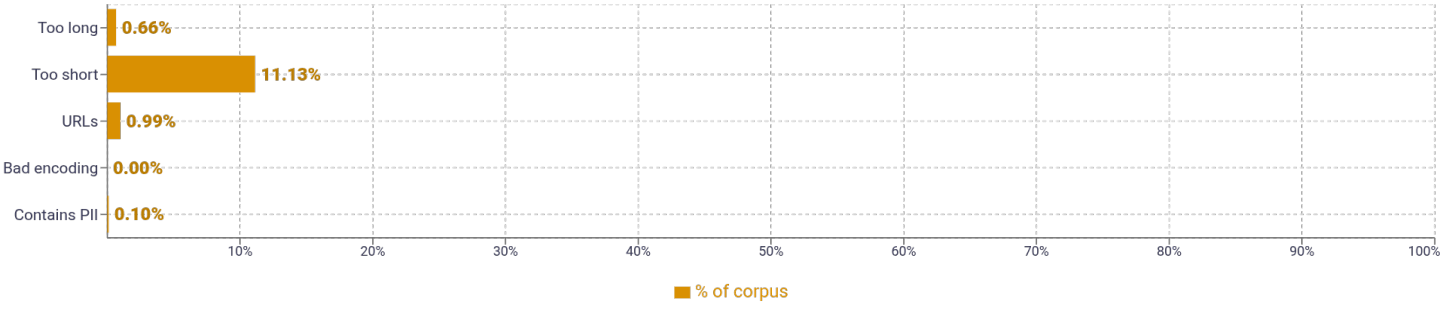


Segment length distribution by token

≤ **49** tokens = **40M** segments | **12M** duplicates
> **50** tokens = **7.9M** segments | **795K** duplicates



Segment noise distribution



Frequent n-grams

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

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				