

General overview

Corpus	Date	Language
prs_Arab.jsonl.tsv	9/21/2024	Dari (prs)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
2,838,831	69,003,437	33,567,402 (48.65 %)	2B	9,501,028,791	15.54 GB

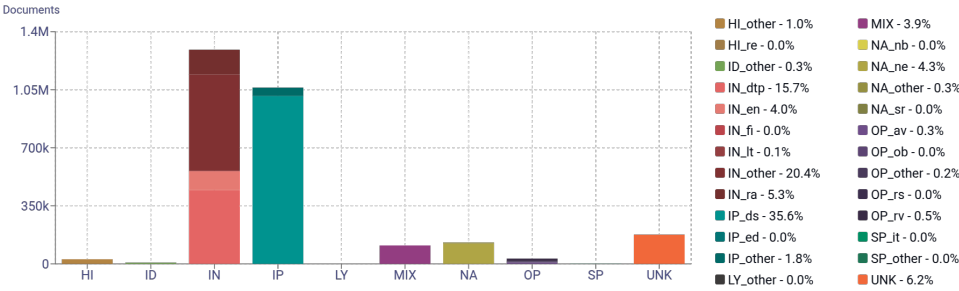
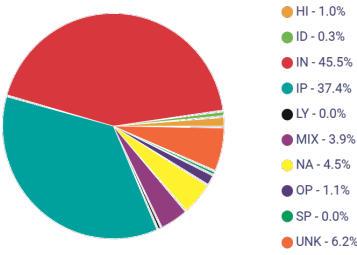
Top 10 domains

Domain	Docs	% of total
wikipedia.org	92K	3.25%
blogfa.com	70K	2.47%
getpaper.ir	39K	1.36%
parscenter.com	35K	1.25%
transline.ir	29K	1.04%
persianblog.ir	21K	0.75%
blogsky.com	18K	0.65%
mihanblog.com	18K	0.64%
p30download.com	14K	0.50%
blog.ir	10K	0.37%

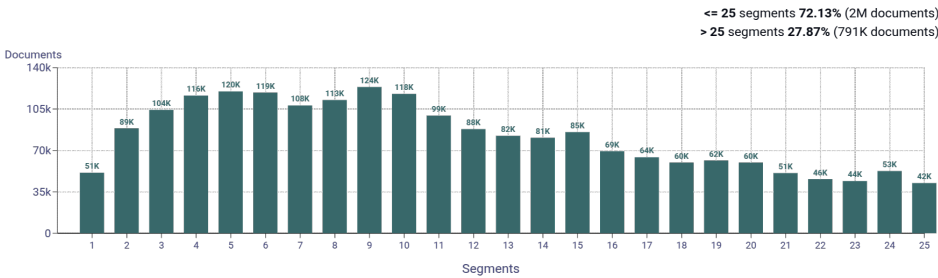
Top 10 TLDs

Domain	Docs	% of total
com	1.3M	45.57%
ir	1.1M	39.71%
org	136K	4.78%
net	66K	2.33%
ac.ir	33K	1.16%
pl	21K	0.76%
de	20K	0.71%
nl	17K	0.59%
co	16K	0.55%
be	14K	0.48%

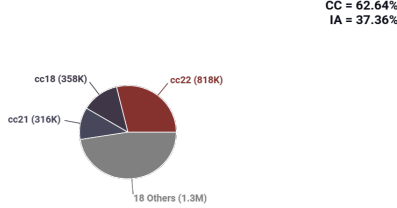
Register labels



Documents size (in segments)

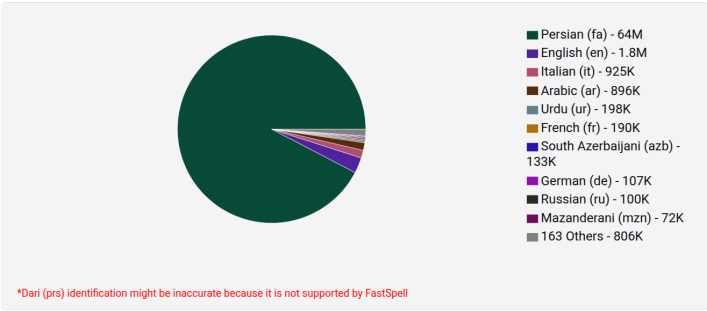


Documents by collection

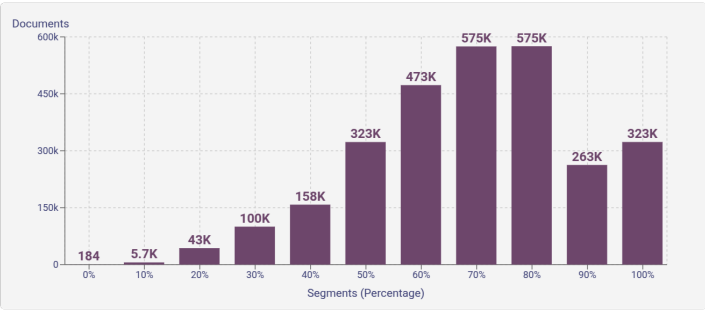


Language Distribution

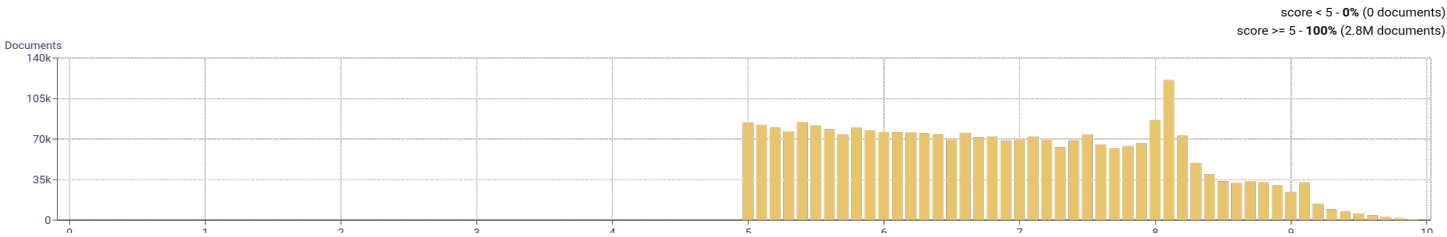
Number of segments in the Dari (prs) corpus



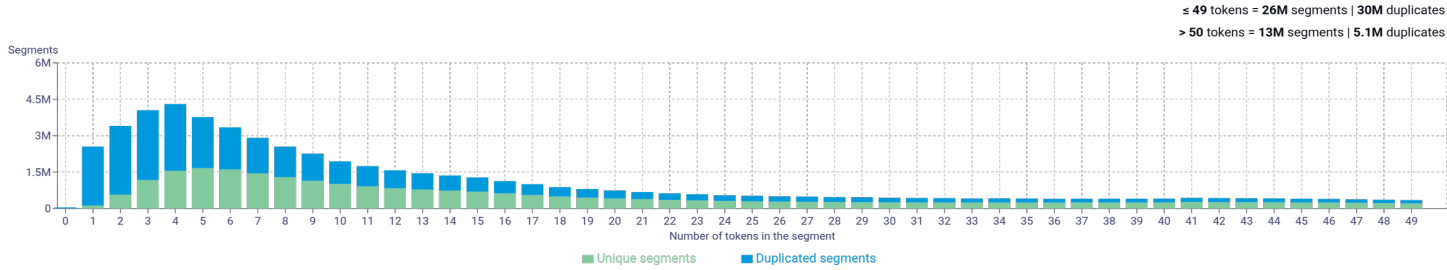
Percentage of segments in Dari (prs) inside documents



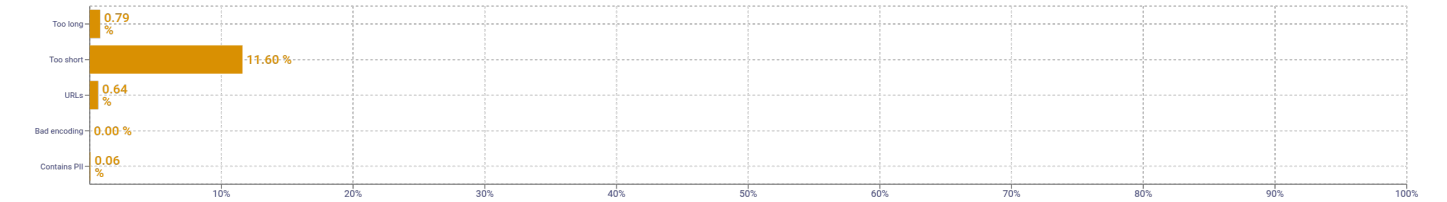
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	سیستم 4353800 تولید 5068136 میشود 5332831 سنگ 5948337 می 18632599
2	فرار می 523116 ماشین آلان 658767 نرم افزار 1465291 می کند 1535242 سنگ شکن 1859803
3	سنگ شکن سنگ 158964 سنگ شکن فکی 164093 شن و ماسه 167642 سنگ شکن مخروطی 173864 تجربه و تحلیل 343046
4	سنگ شکن برای فروش 38237 متن ساختگی با تولید 38450 نصب و راه اندازی 40699 سنگ شکن صریح ای 49525 دستگاه های سنگ شکن 86516
5	سازگی نامفهوم از صنعت چاپ 36005 تولید سازگی نامفهوم از صنعت 36213 ساختگی با تولید سازگی نامفهوم 37284 متن ساختگی با تولید سازگی 38011 ایپسوم متن ساختگی با تولید 38282

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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>, , , 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/pablopt6n/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				