HPLT Analytics report

General overview

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

Volumes

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

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

Documents size (in segments)

<= 25 segments **72.13%** (2M documents)
> 25 segments **27.87%** (791K documents)
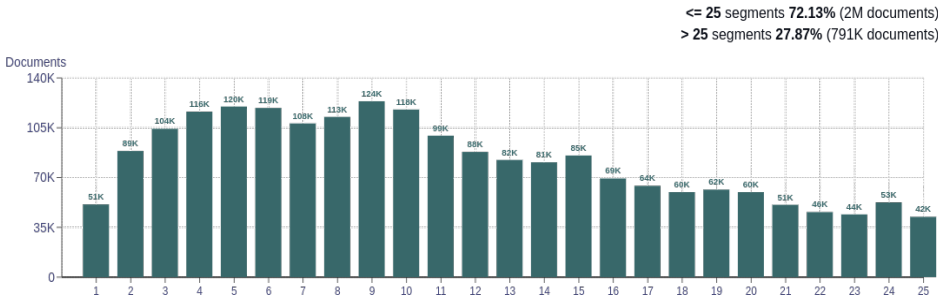
Documents by collection

cc18 (358K), cc22 (818K), cc21 (316K), 18 Others (1.3M)
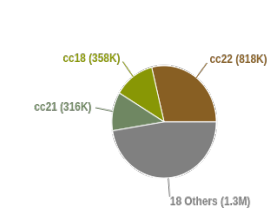
Language Distribution

Number of segments

- Persian (fa) - 64M
- English (en) - 1.8M
- Italian (it) - 925K
- Arabic (ar) - 896K
- Urdu (ur) - 198K
- French (fr) - 190K
- South Azerbaijani (azb) - 133K
- German (de) - 107K
- Russian (ru) - 100K
- Mazanderani (mzn) - 72K
- 163 Others - 806K

*Dari (prs) identification might be inaccurate because language is not supported by Fasttext
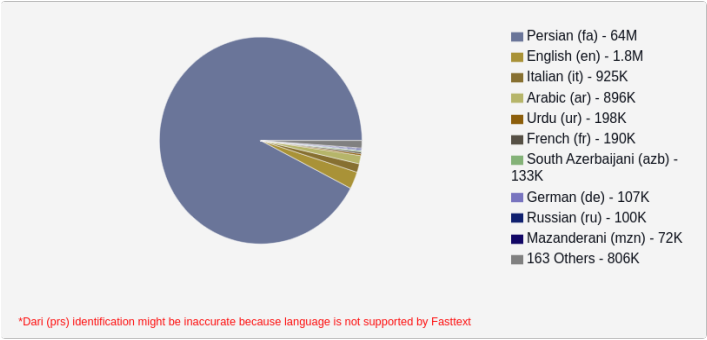
Percentage of segments in Dari (prs) inside documents
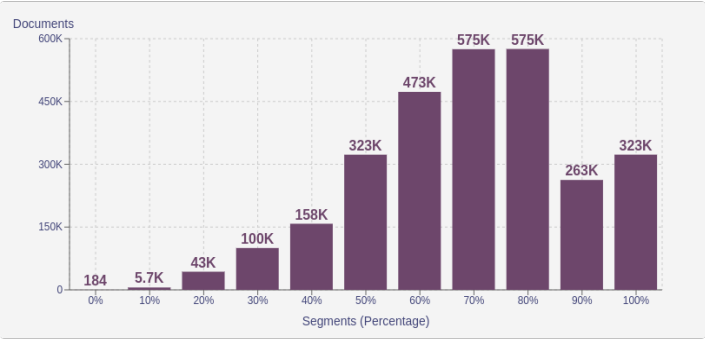
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (2.8M documents)

Segment length distribution by token

<= 49 tokens = **26M** segments | **30M** duplicates
> 50 tokens = **13M** segments | **5.1M** duplicates

- Unique segments
- Duplicated segments

Segment noise distribution

Too long	0.79 %
Too short	11.60 %
URLs	0.64 %
Bad encoding	0.00 %
Contains PII	0.06 %
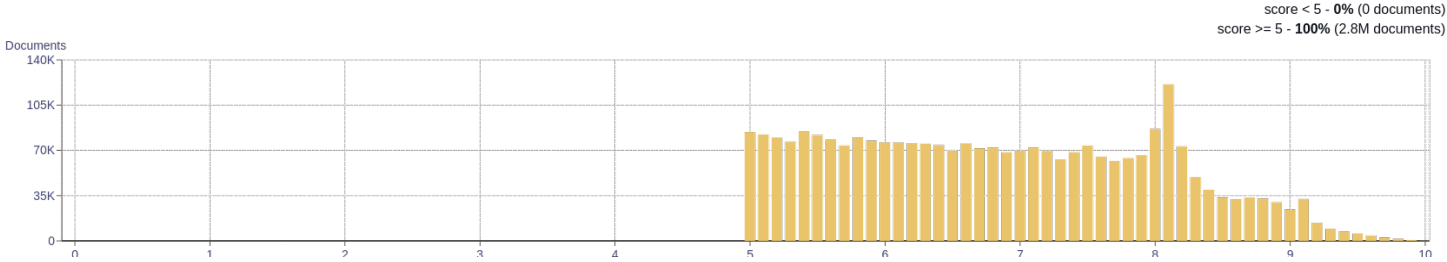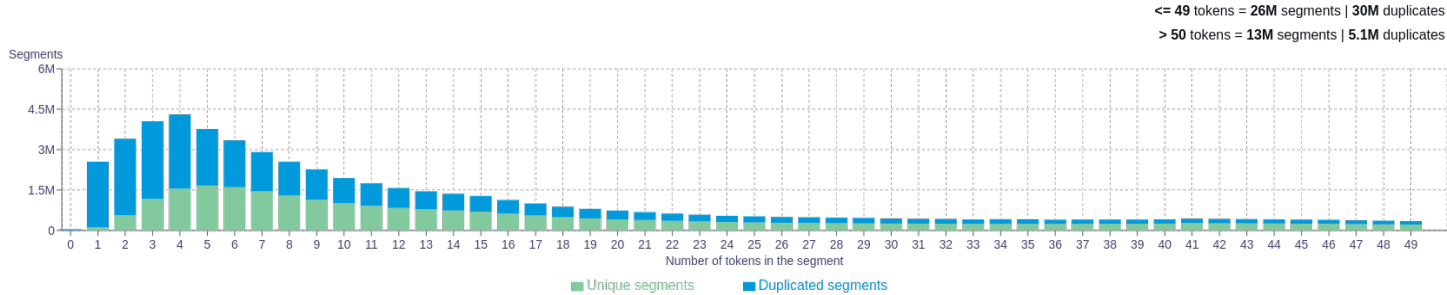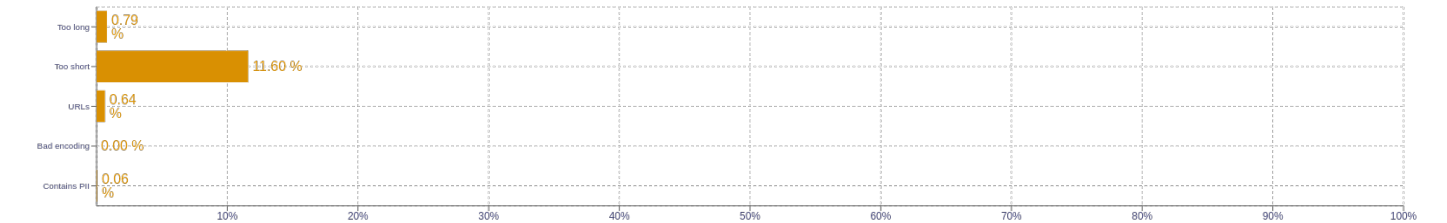
Frequent n-grams

Size	n-grams
1	می | 18632599 سنگ | 5948337 میشود | 5332831 تولید | 5068136 سیستم | 4353800
2	فرار می | 523116 ماشین آلات | 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 | 173064 تجزیه و تحلیل | 343046
4	سنگ شکن برای فروش | 38237 متن ساختگی با تولید | 38450 نصب و راه اندازی | 40699 سنگ شکن ضربه ای | 49525 دستگاه های سنگ شکن | 86516
5	تولید سادگی نامفهوم از صنعت | 36213 ساختگی با تولید سادگی نامفهوم | 37284 متن ساختگی با تولید سادگی | 38011 ایبسوم متن ساختگی با تولید | 38282 سادگی نامفهوم از صنعت چاپ | 36005

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