

General overview

Corpus	Analytics date	Language
azb_Arab.jsonl.tsv	9/27/2024	South Azerbaijani (azb)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
66,112	2,389,200	1,055,552 (44.18 %)	49M	446.12 MB	257,866,139

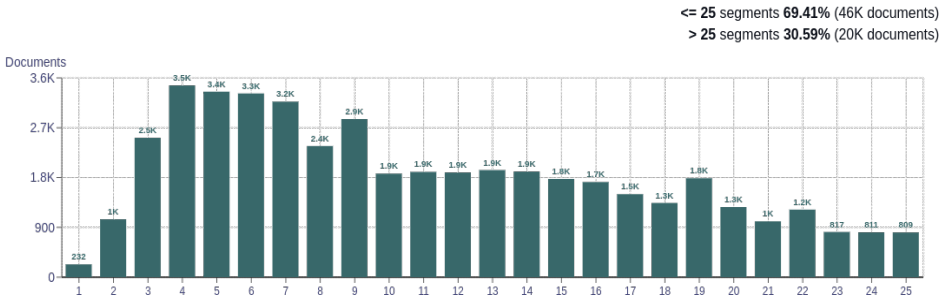
Top 10 domains

Domain	Docs	% of total
blogfa.com	14K	21.68
wikipedia.org	11K	16.86
axar.az	4.7K	7.09
trt.net.tr	4.5K	6.80
arzublog.com	2.9K	4.45
ishiq.net	2.9K	4.35
baybak.com	2.1K	3.12
bilimsesi.com	1.5K	2.23
blogsky.com	1.4K	2.12
mihanblog.com	1.2K	1.88

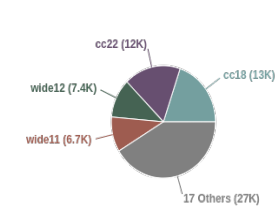
Top 10 TLDs

Domain	Docs	% of total
com	33K	50.20
org	13K	19.92
ir	5K	7.60
az	4.7K	7.09
net.tr	4.5K	6.80
net	3.3K	4.96
info	380	0.57
biz	271	0.41
se	253	0.38
ca	209	0.32

Documents size (in segments)

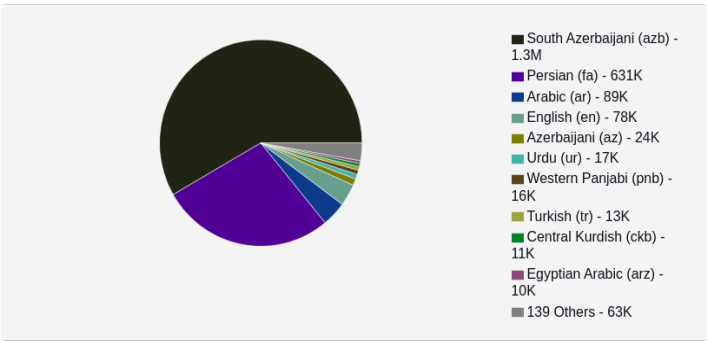


Documents by collection

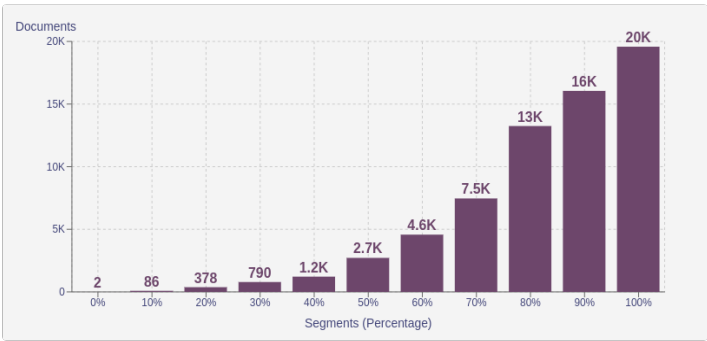


Language Distribution

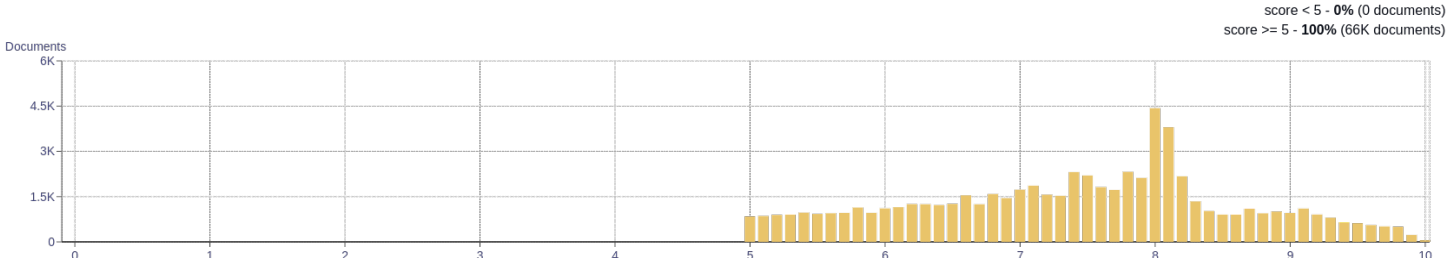
Number of segments



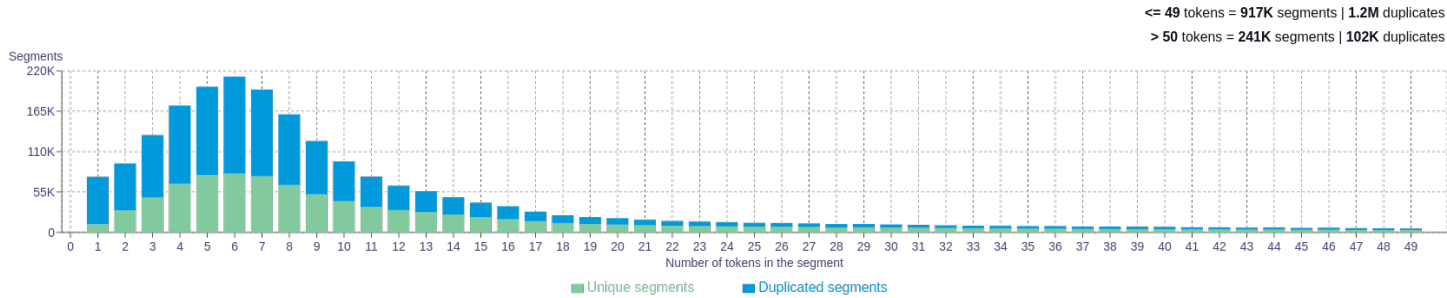
Percentage of segments in South Azerbaijani (azb) inside documents



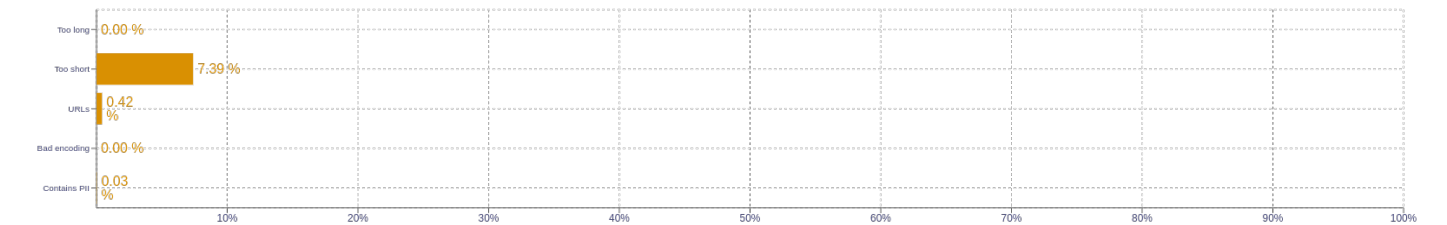
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>172533 ایله 156791 دا 109314 کیمی 98960 جوخ 92750 دیر</div>
2	<div>14647 جی ایلده read more 9442 6535 بونا گۆره 6317 داها جوخ 5500 داها آرتیق</div>
3	<div>3082 گۆره داها آرتیق 3079 داها آرتیق بیلگیلر 3078 آرتیق بیلگیلر تا یا بیلرسینیز 2740 ایشلدنلری طرفیندن یارانمیش 2718 ویکیپدیا سینین ایشلدنلری طرفیندن</div>
4	<div>3078 گۆره داها آرتیق بیلگیلر 3078 داها آرتیق بیلگیلر تا یا بیلرسینیز 2718 ویکیپدیا سینین ایشلدنلری طرفیندن یارانمیش 2099 اینگیلیسجه ویکیپدیا سینین ایشلدنلری طرفیندن</div> <div>1537 ایران مالیکی محروسه سینده</div>
5	<div>3078 گۆره داها آرتیق بیلگیلر تا یا بیلرسینیز 2099 اینگیلیسجه ویکیپدیا سینین ایشلدنلری طرفیندن یارانمیش 561 بیشیک موبایل ویی ایلده د 496 ی نیمی ترکیمینده جیخیش آندیب</div> <div>414 ایلام ایلام ایلام ایلام ایلام</div>

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.stinfo.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/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>