HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-docslite.az.tsv	6/7/2024	Azerbaijani (az)

Volumes

Docs	Segments	Unique segments	Tokens	Size
1,097,781	139,346,876	48,887 (0.04 %)	1.5B	8.99 GB

Top 10 domains

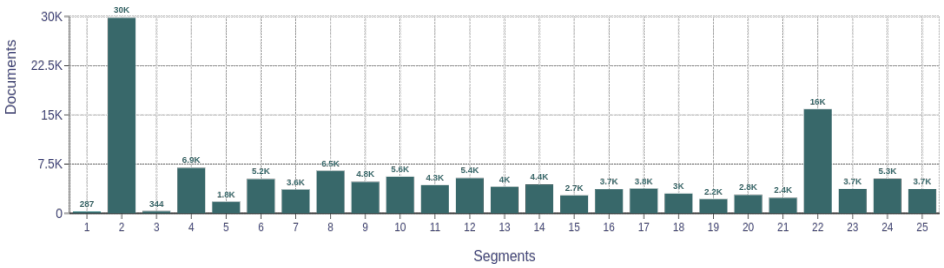
Domain	Docs	% of total
beyaz.az	37K	3.39
wikipedia.org	16K	1.49
ictnews.az	16K	1.43
haqqinda.az	11K	1.05
cumhuriyyet.net	11K	1.00
deyerler.org	9.5K	0.86
faktxeber.com	8.8K	0.80
gununsesi.info	8.6K	0.78
bul.az	8.4K	0.76
rublika.az	6.7K	0.61

Top 10 TLDs

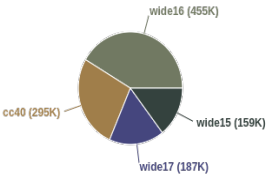
Domain	Docs	% of total
az	635K	57.85
com	175K	15.92
org	78K	7.08
info	44K	3.99
net	42K	3.82
gov.az	29K	2.61
edu.az	15K	1.35
tv	12K	1.10
ru	12K	1.07
biz	8.2K	0.74

Documents size (in segments)

<= 25 segments **12.05%** (132K documents)
> 25 segments **87.95%** (964K documents)

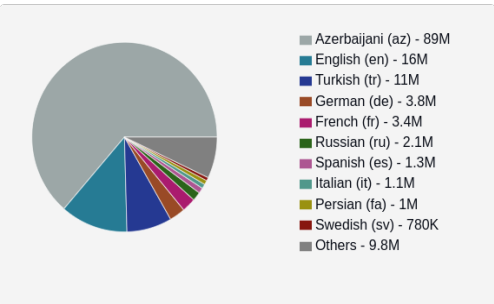


Documents by collection

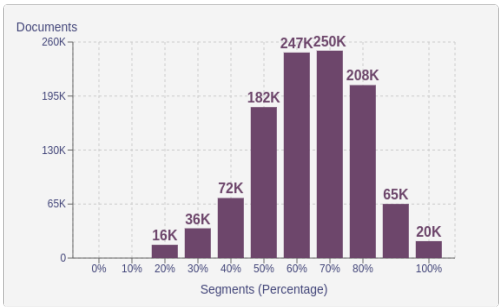


Language Distribution

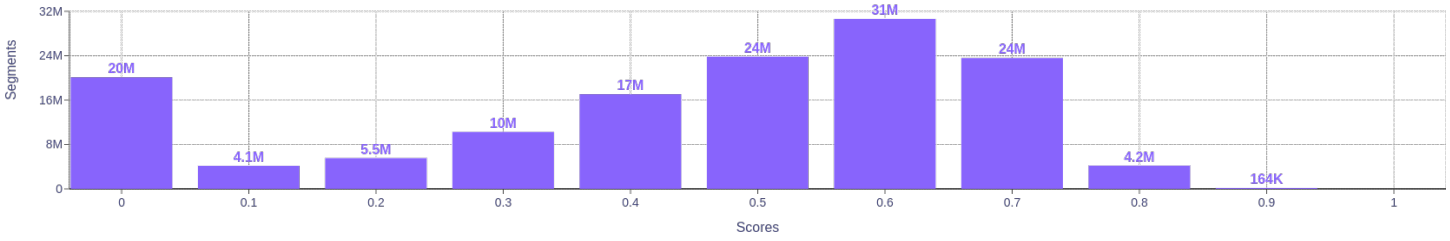
Number of segments



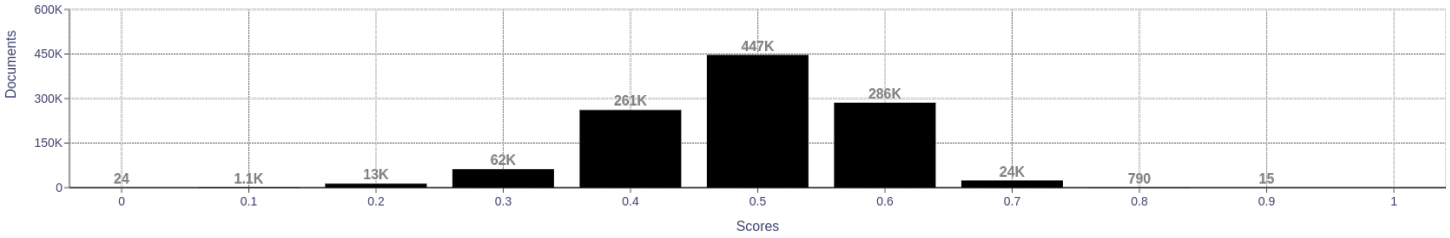
Percentage of segments in Azerbaijani (az) inside documents



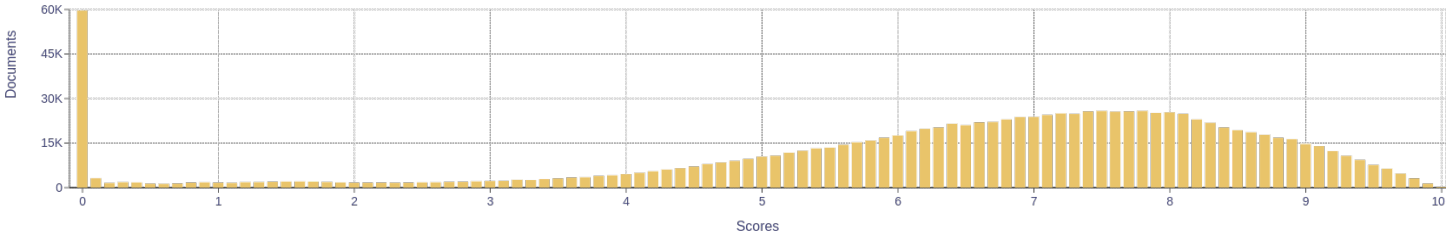
Distribution of segments by fluency score



Distribution of documents by average fluency score



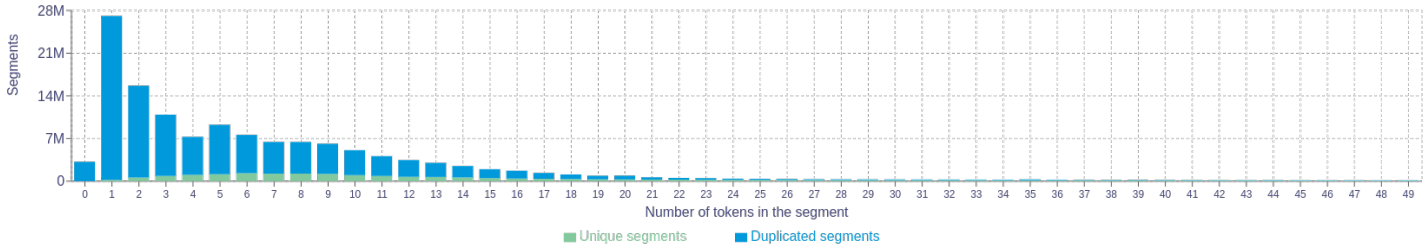
Distribution of documents by document score



Segment length distribution by token

<= 49 tokens = **20M** segments | **115M** duplicates

> 50 tokens = **5.1M** segments | **1.4M** duplicates



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