HPLT Analytics report

General overview

Corpus	Analytics date	Language
HPLT-v2-azj_Latn.tsv	9/28/2024	Azerbaijani (azj)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
6,484,902	126,612,778			21.06 GB	19,502,117,097

Top 10 domains

Domain	Docs	% of total
azadliq.org	196K	3.03
wikipedia.org	187K	2.88
publika.az	145K	2.24
report.az	123K	1.90
amerikaninsesi.org	108K	1.67
trend.az	99K	1.53
ictnews.az	97K	1.50
stadium.az	95K	1.47
metbuat.az	77K	1.19
haqqinda.az	68K	1.04

Top 10 TLDs

Domain	Docs	% of total
az	3.8M	59.10
com	985K	15.19
org	686K	10.57
info	192K	2.96
net	177K	2.73
gov.az	177K	2.72
edu.az	67K	1.03
tv	63K	0.97
biz	53K	0.81
ws	26K	0.41

Documents size (in segments)

<= 25 segments **82.24%** (5.3M documents)
> 25 segments **17.76%** (1.2M documents)
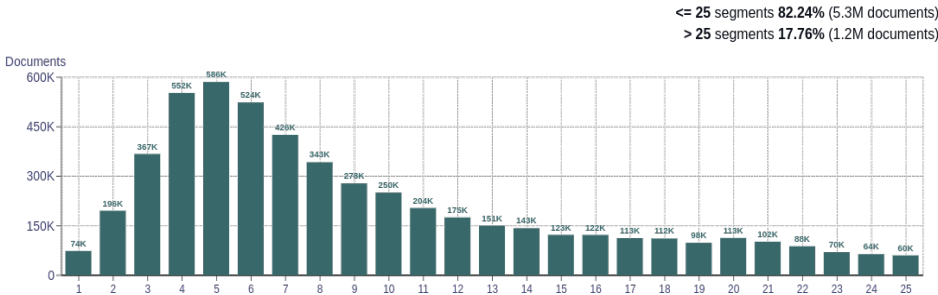
Documents by collection

cc18 (1.1M)
cc17 (663K)
cc22 (1.2M)
18 Others (3.5M)
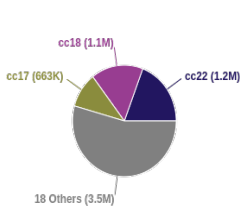
Language Distribution

Number of segments

- Azerbaijani (az) - 105M
- Turkish (tr) - 5.8M
- Italian (it) - 4.3M
- English (en) - 4.2M
- German (de) - 1.6M
- Russian (ru) - 720K
- French (fr) - 627K
- Norwegian Bokmål (nb) - 355K
- Spanish (es) - 288K
- Chinese (zh) - 286K
- 164 Others - 3.4M

*Azerbaijani (azj) identification might be inaccurate because language is not supported by Fasttext
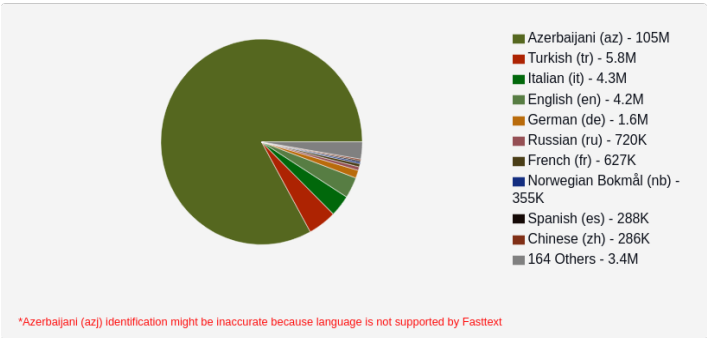
Percentage of segments in Azerbaijani (azj) inside documents
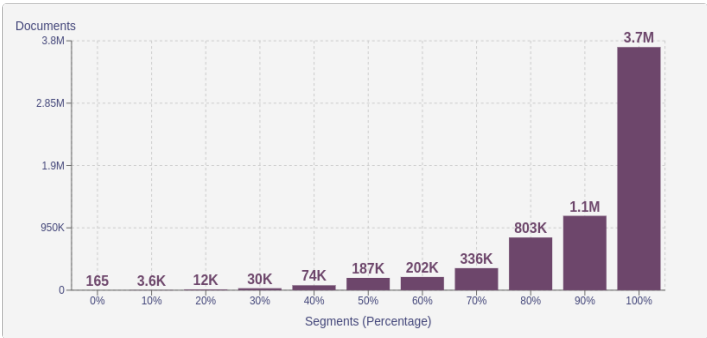
0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
165	3.6K	12K	30K	74K	187K	202K	336K	803K	1.1M	3.7M

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (6.5M documents)

Segment noise distribution

Category	Percentage
Too long	0.00 %
Too short	14.64 %
URLs	2.61 %
Bad encoding	0.00 %
Contains PII	0.10 %
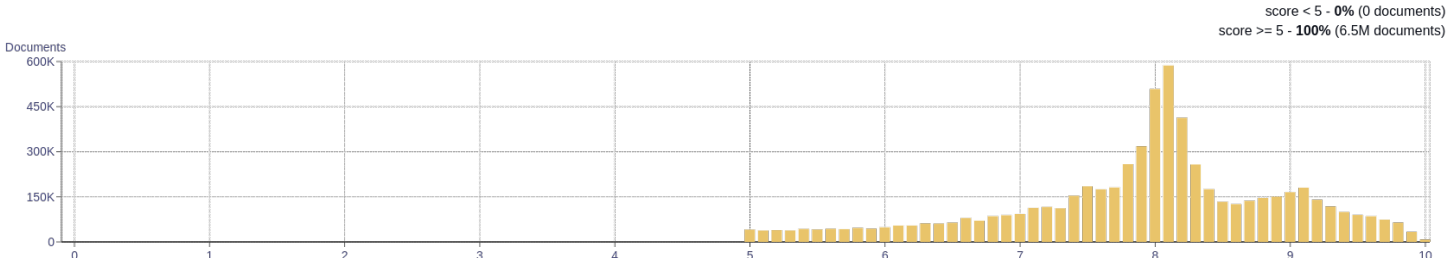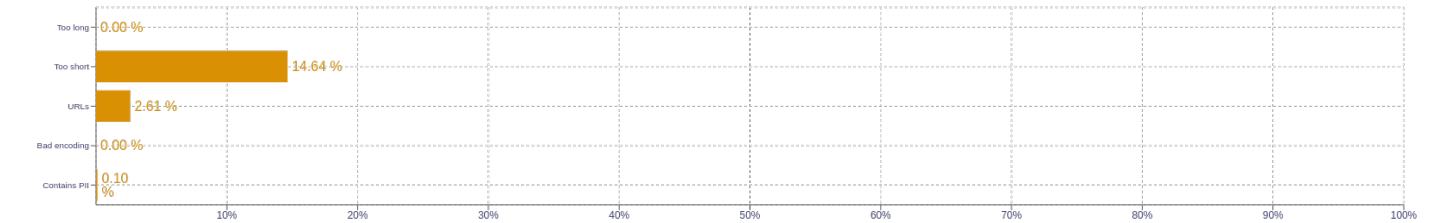
About HPLT Analytics

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