HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
azj_Latn.jsonl.tsv	6/5/2025	Azerbaijani (azj)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
6,484,902	126,568,581	50,737,488 (40.09 %)	3.2B	19,502,117,097	21.06 GB

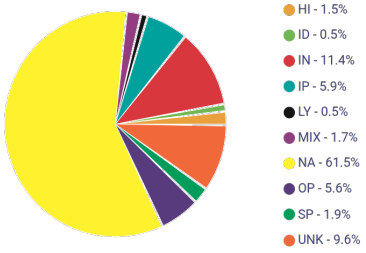
Top 10 domains

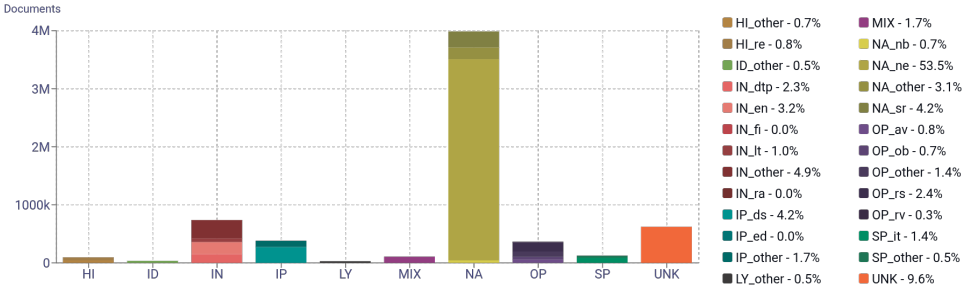
Domain	Docs	% of total
azadliq.org	196K	3.03%
wikipedia.org	187K	2.88%
publika.az	145K	2.24%
report.az	123K	1.90%
amerikaninsesi.org	108K	1.67%
trend.az	99K	1.53%
ictnews.az	97K	1.50%
stadium.az	95K	1.47%
metbuat.az	77K	1.19%
haqqinda.az	68K	1.04%

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%

Register labels

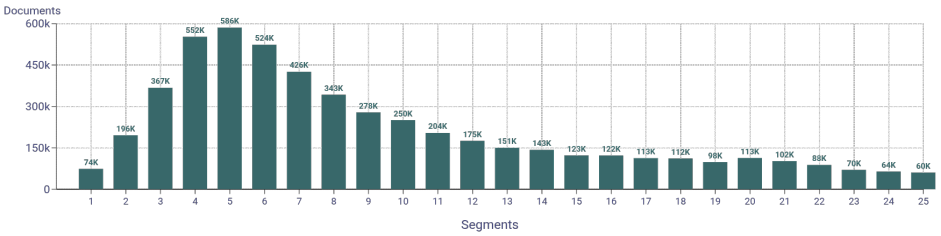




MT:4.6% | 299K Documents

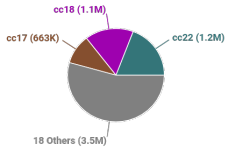
Documents size (in segments)

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



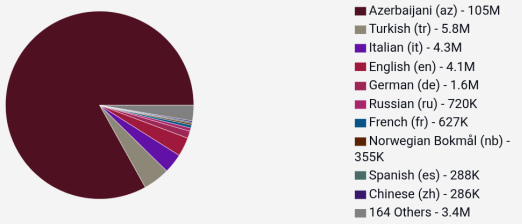
Documents by collection

CC = 57.38%
IA = 42.62%



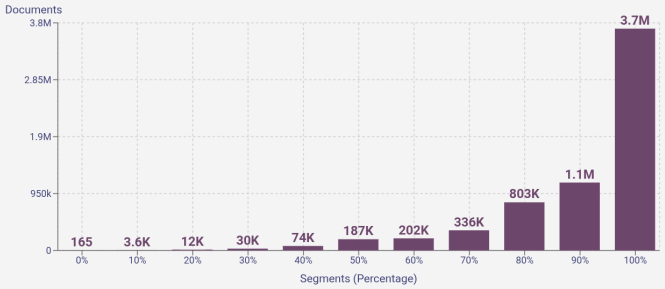
Language Distribution

Number of segments in the Azerbaijani (azj) corpus



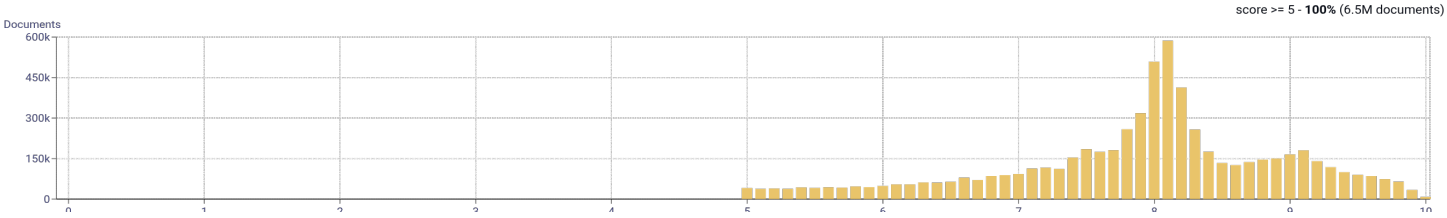
*Azerbaijani (azj) identification might be inaccurate because it is not supported by FastSpell

Percentage of segments in Azerbaijani (azj) inside documents



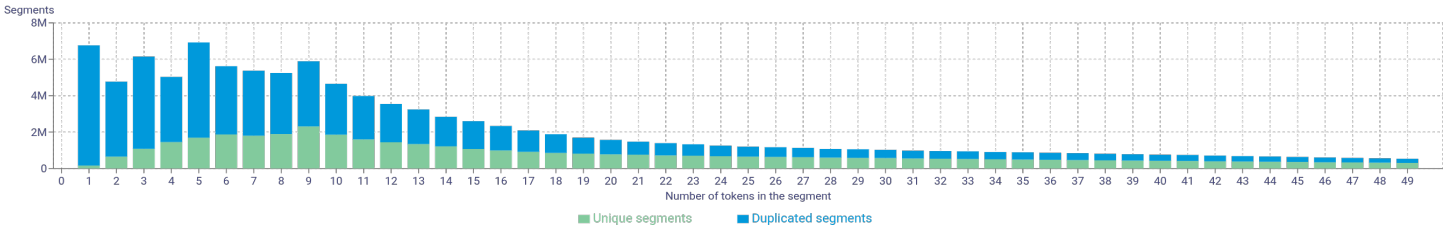
Distribution of documents by document score

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

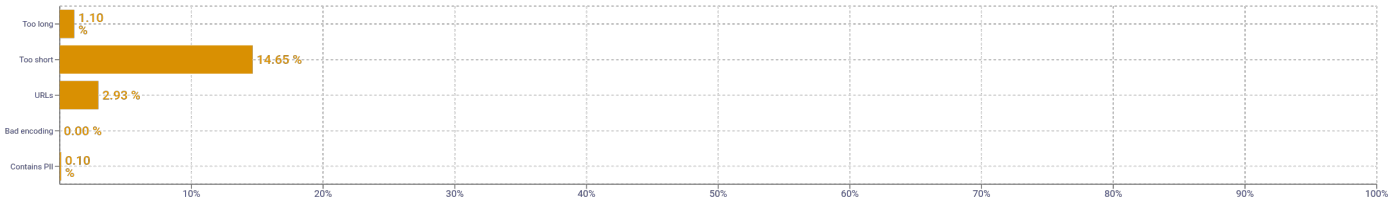


Segment length distribution by token

≤ 49 tokens = **40M** segments | **68M** duplicates
> 50 tokens = **18M** segments | **7.5M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	azərbaycan | 11656864 · dövlət | 5207929 · baş | 5090151 · yeni | 4655229 · edib | 4437511
2	azərbaycan respublikasının | 1660078 · xəbər verir | 1546431 · azərbaycan respublikası | 1262587 · dəfə oxunub | 951651 · qeyd edək | 906912
3	istinadən xəbər verir | 367922 · prezident ilham əliyev | 319924 · azərbaycan respublikasının prezidenti | 313585 · azərbaycan respublikası prezidentinin | 250507 · prezidenti ilham əliyev | 179886
4	azərbaycan respublikasının prezidenti ilham | 112382 · ekologiya və təbii sərvətlər | 100004 · əmək və əhalinin sosial | 93227 · azərbaycan prezidenti ilham əliyev | 89811 · ümummilli lider heydər əliyevin | 86444
5	əmək və əhalinin sosial müdafiəsi | 90322 · azərbaycan respublikasının prezidenti ilham əliyev | 73882 · ekologiya və təbii sərvətlər nazirliyinin | 54883 · azərbaycan respublikasının prezidenti cənab ilham | 38849 · bərədə azərbaycan respublikası qanununun layihəsi | 37182

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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av