HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-tgk_Cyrl	9/18/2025	Tajik

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
2,574,512	41,842,368	30,508,814 (72.91 %)	1.3B	7,908,927,830	13.4 GB

Top 10 domains

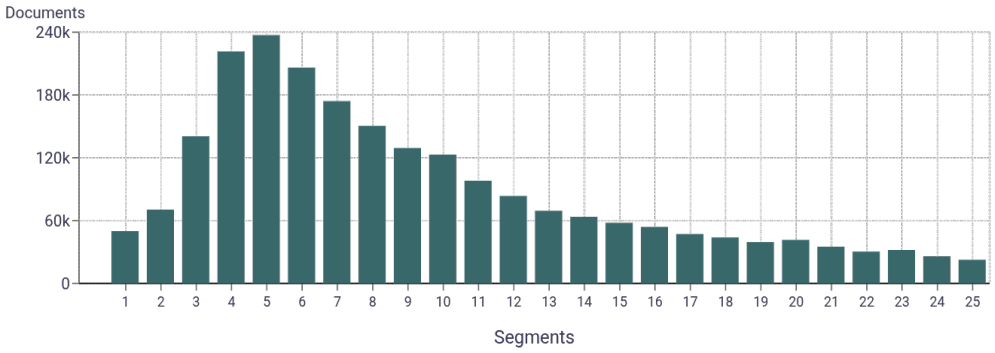
Domain	Docs	% of total
ozodi.org	155K	6.02%
kun.uz	100K	3.89%
ozodlik.org	83K	3.23%
sputnik-tj.com	64K	2.48%
daryo.uz	45K	1.75%
sputniknews-uz.com	39K	1.51%
sports.uz	38K	1.50%
xs.uz	37K	1.45%
islom.uz	35K	1.35%
xabar.uz	31K	1.22%

Top 10 TLDs

Domain	Docs	% of total
uz	1.2M	45.86%
com	408K	15.87%
tj	383K	14.86%
org	333K	12.93%
info	63K	2.45%
kz	35K	1.38%
net	29K	1.13%
asia	26K	1.01%
ru	24K	0.93%
tv	10K	0.39%

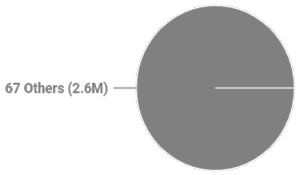
Documents size (in segments)

≤ 25 segments **87.3%** (2.2M documents)
> 25 segments **12.7%** (327K documents)



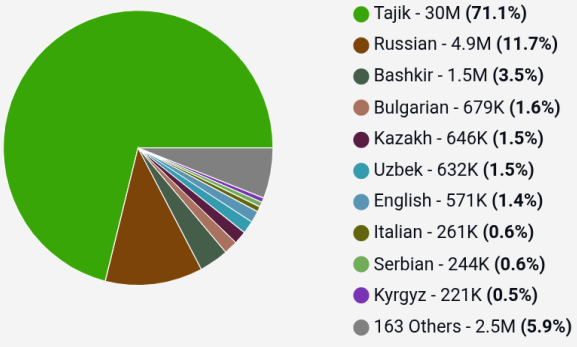
Document collections

CC = 94.57%
IA = 5.43%



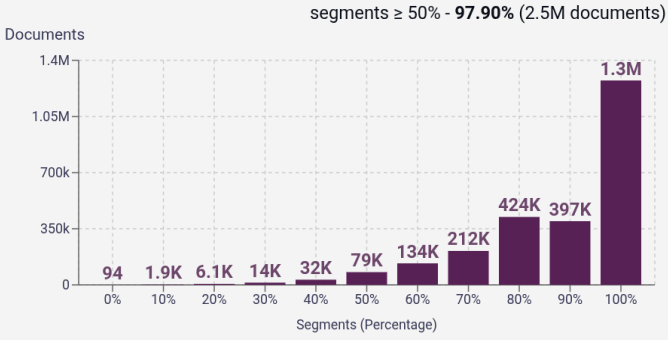
Language Distribution

Number of segments in the Tajik corpus



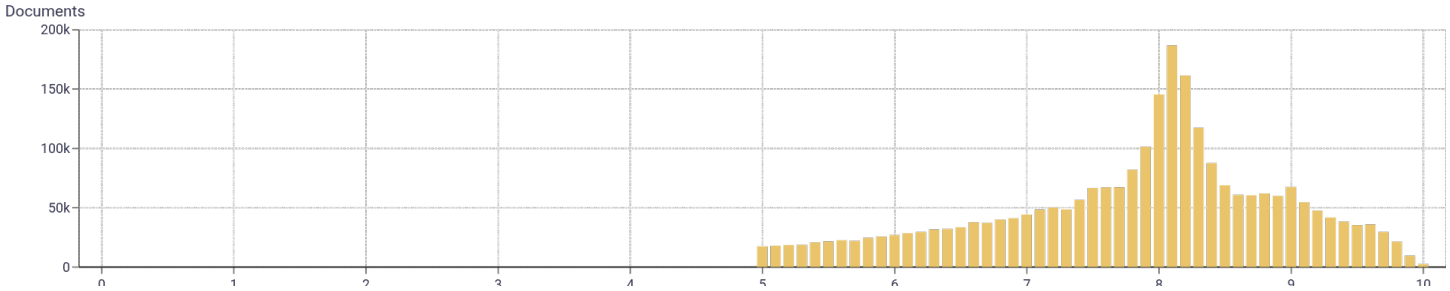
Percentage of segments in Tajik inside documents

segments < 50% - **2.10%** (54K documents)
segments ≥ 50% - **97.90%** (2.5M documents)



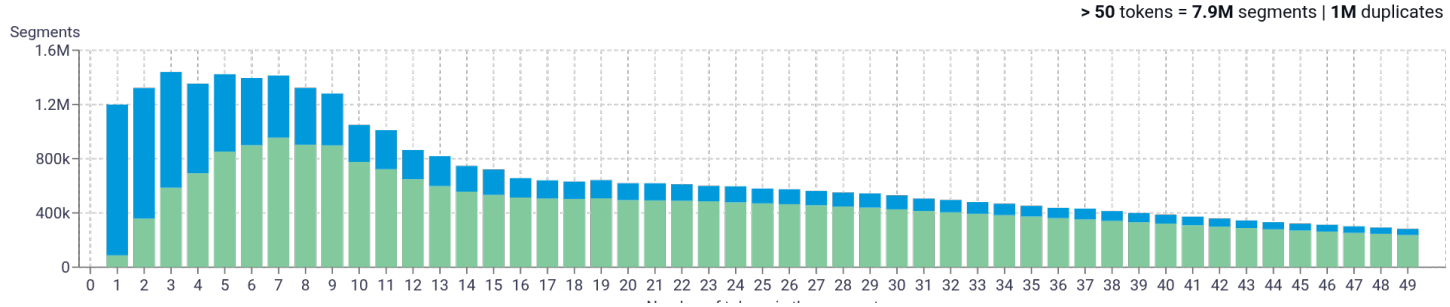
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (2.6M documents)

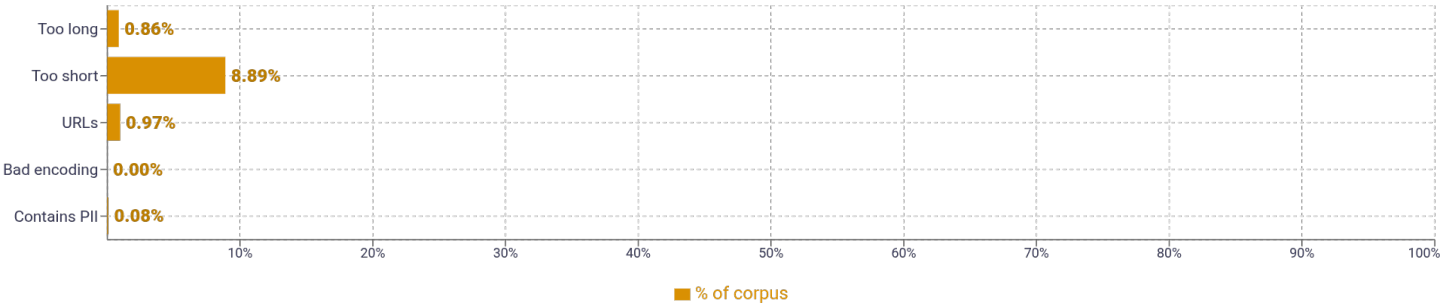


Segment length distribution by token

≤ 49 tokens = **34M** segments | **10M** duplicates
> 50 tokens = **7.9M** segments | **1M** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	билан | 5,410,438 · бу | 3,832,221 · учун | 3,640,892 · бир | 3,208,216 · тоҷикистон | 2,772,940
2	ҷумҳурии тоҷикистон | 1,055,368 · ӯзбекистон республикаси | 672,163 · эмомалӣ раҳмон | 358,238 · бу ҳақда | 333,299 · президенти ҷумҳурии | 263,119
3	президенти ҷумҳурии тоҷикистон | 239,686 · муҳтарам эмомалӣ раҳмон | 194,552 · ҳукумати ҷумҳурии тоҷикистон | 131,611 · сулҳу ваҳдати миллӣ | 116,286 · асосгузори сулҳу ваҳдати | 114,531
4	асосгузори сулҳу ваҳдати миллӣ | 112,890 · тоҷикистон муҳтарам эмомалӣ раҳмон | 99,842 · ҷумҳурии тоҷикистон муҳтарам эмомалӣ | 98,967 · президенти ҷумҳурии тоҷикистон муҳтарам | 98,448 · маҷлиси олии ҷумҳурии тоҷикистон | 60,583
5	президенти ҷумҳурии тоҷикистон муҳтарам эмомалӣ | 97,516 · ҷумҳурии тоҷикистон муҳтарам эмомалӣ раҳмон | 95,979 · қолдириш учун сайтда рӯйхатдан ӯтинг | 31,252 · изоҳ қолдириш учун сайтда рӯйхатдан | 31,252 · маҷлиси намояндагони маҷлиси олии ҷумҳурии | 30,013

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.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	**HI**	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	**Informational persuasion**	**IP**	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	**Opinion**	**OP**
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	**Informational description**	**IN**	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				