HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
tgk_Cyrl.jsonl.tsv	9/16/2024	Tajik (tg)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,261,259	24,851,003	14,469,071 (58.22 %)	770M	4,565,968,108	7.75 GB

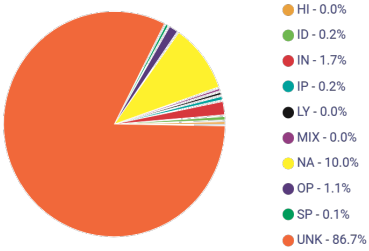
Top 10 domains

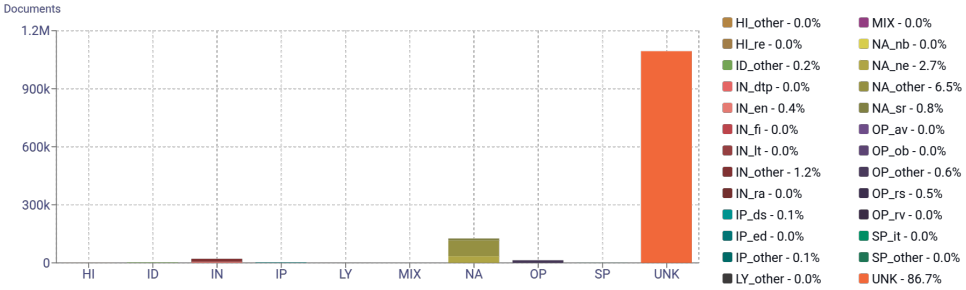
Domain	Docs	% of total
ozodi.org	123K	9.77%
ozodlik.org	80K	6.36%
kun.uz	52K	4.13%
wikipedia.org	25K	1.97%
ozodagon.com	21K	1.67%
khovar.tj	18K	1.40%
islom.uz	16K	1.25%
fikr.uz	15K	1.18%
qalampir.uz	15K	1.17%
daryo.uz	13K	0.99%

Top 10 TLDs

Domain	Docs	% of total
uz	501K	39.72%
org	267K	21.19%
com	194K	15.41%
tj	188K	14.90%
info	21K	1.66%
ru	19K	1.52%
net	13K	1.06%
mobi	12K	0.97%
asia	9.9K	0.79%
kz	4.8K	0.38%

Register labels

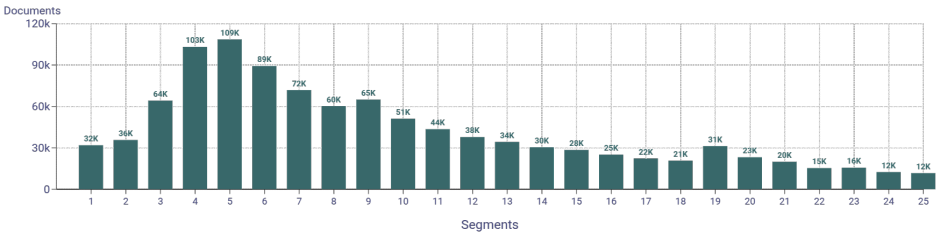




MT:74.8% | 944K Documents

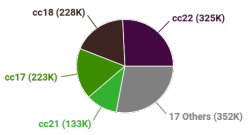
Documents size (in segments)

<= 25 segments **83.48%** (1.1M documents)
> 25 segments **16.52%** (208K documents)



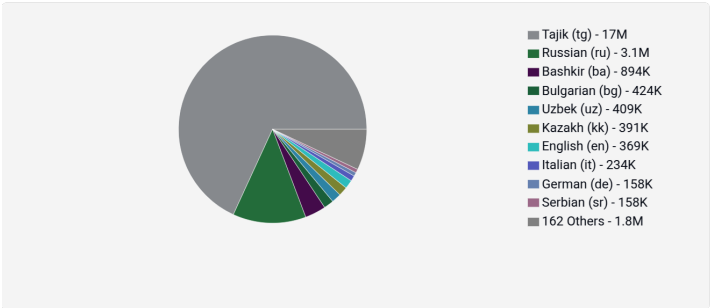
Documents by collection

CC = 75.78%
IA = 24.22%

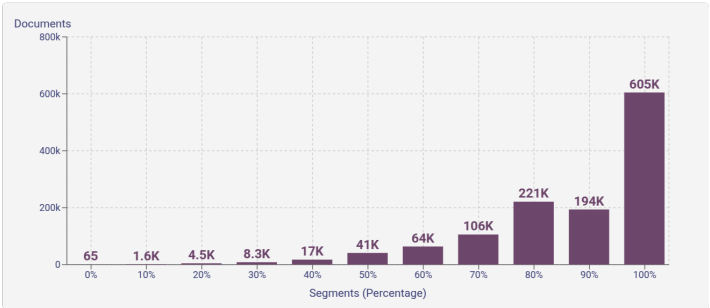


Language Distribution

Number of segments in the Tajik (tg) corpus

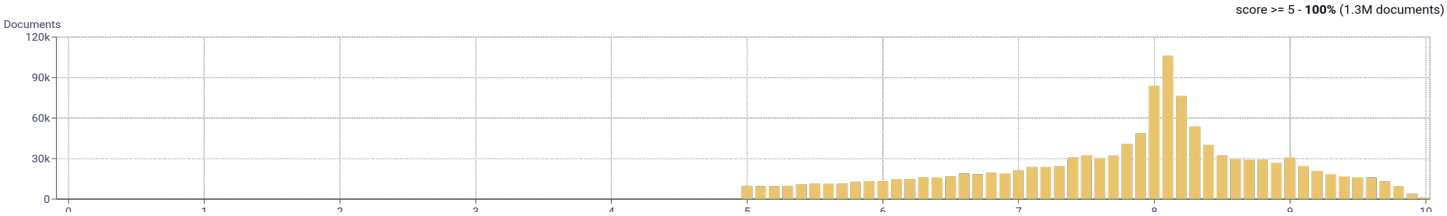


Percentage of segments in Tajik (tg) inside documents



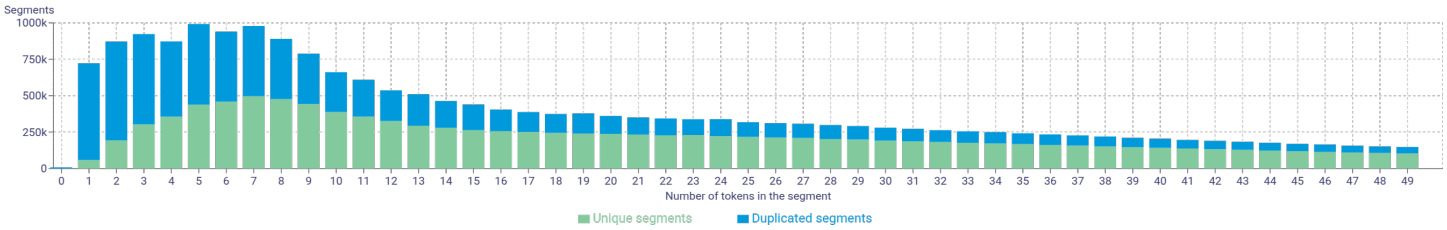
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (1.3M documents)

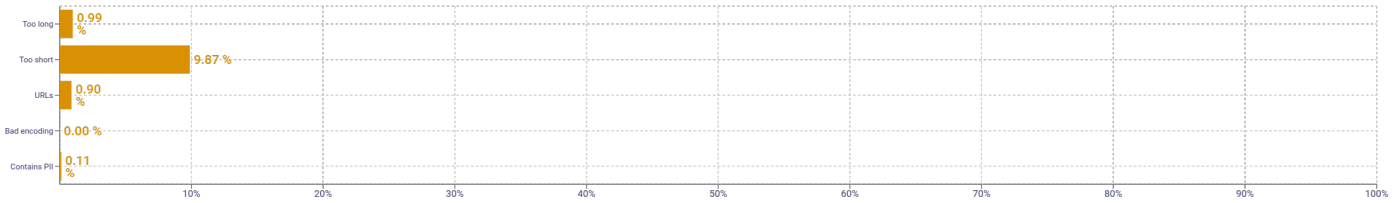


Segment length distribution by token

≤ 49 tokens = **11M** segments | **9M** duplicates
> 50 tokens = **4.7M** segments | **1.4M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	билан | 3299836 · бу | 2456944 · бир | 2146323 · учун | 2134996 · тоҷикистон | 1655856
2	ҷумҳурии тоҷикистон | 661595 · ӯзбекистон республикаси | 603614 · эмомалӣ раҳмон | 170101 · шаҳри душанбе | 159888 · соллаллоҳу алайҳи | 147219
3	президенти ҷумҳурии тоҷикистон | 124169 · вайп вайп вайп | 82986 · муҳтарам эмомалӣ раҳмон | 75996 · ӯзбекистон республикаси олий | 74878 · ҳукумати ҷумҳурии тоҷикистон | 74181
4	вайп вайп вайп вайп | 82660 · асосгузори сулҳу ваҳдати миллӣ | 46294 · ӯзбекистон республикаси вазирлар маҳкамасининг | 46155 · ӯзбекистон республикаси олий мажлиси | 41447 · соллаллоҳу алайҳи ва саллам | 41224
5	вайп вайп вайп вайп вайп | 82394 · саги пидорасу саги пидорасу саги | 39003 · пидорасу саги пидорасу саги пидорасу | 38904 · президенти ҷумҳурии тоҷикистон муҳтарам эмомалӣ | 36225 · ҷумҳурии тоҷикистон муҳтарам эмомалӣ раҳмон | 35809

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