HPLT Analytics report

General overview

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

Volumes

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

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

Documents size (in segments)

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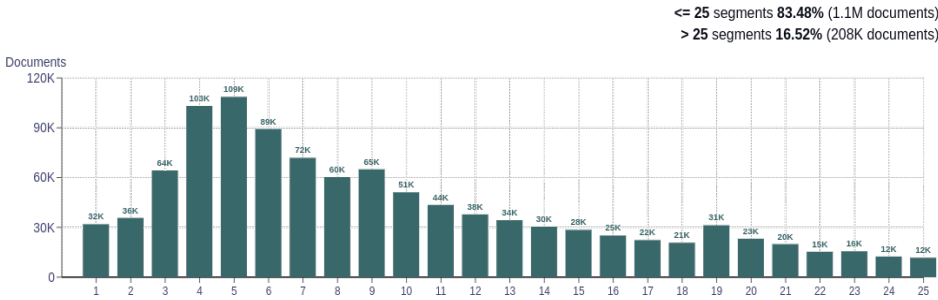
Documents by collection

cc18 (228K), cc22 (325K), cc17 (223K), cc21 (133K), 17 Others (352K)
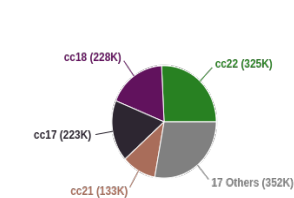
Language Distribution

Number of segments

- Tajik (tg) - 17M
- Russian (ru) - 3.1M
- Bashkir (ba) - 894K
- Bulgarian (bg) - 424K
- Uzbek (uz) - 409K
- Kazakh (kk) - 391K
- English (en) - 369K
- Italian (it) - 234K
- German (de) - 158K
- Serbian (sr) - 158K
- 162 Others - 1.8M
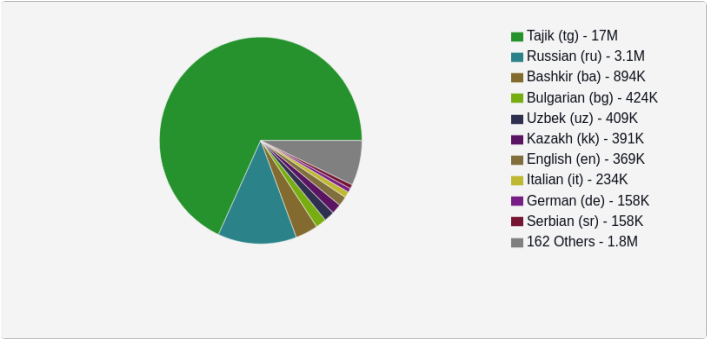
Percentage of segments in Tajik (tg) inside documents
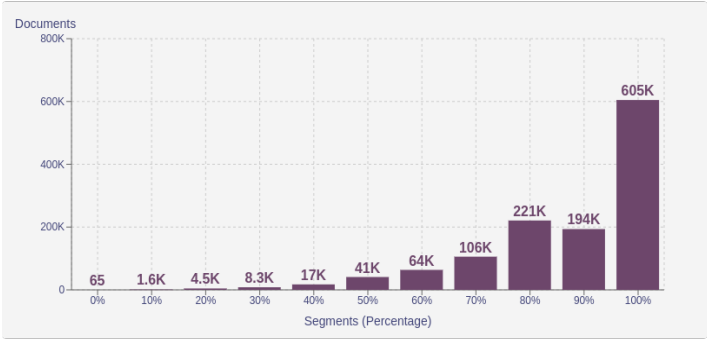
Segments (Percentage)	Documents
0%	65
10%	1.6K
20%	4.5K
30%	8.3K
40%	17K
50%	41K
60%	64K
70%	106K
80%	221K
90%	194K
100%	605K

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

Number of tokens in the segment

- Unique segments
- Duplicated segments

Segment noise distribution

Too long	0.99 %
Too short	9.87 %
URLs	0.90 %
Bad encoding	0.00 %
Contains PII	0.11 %
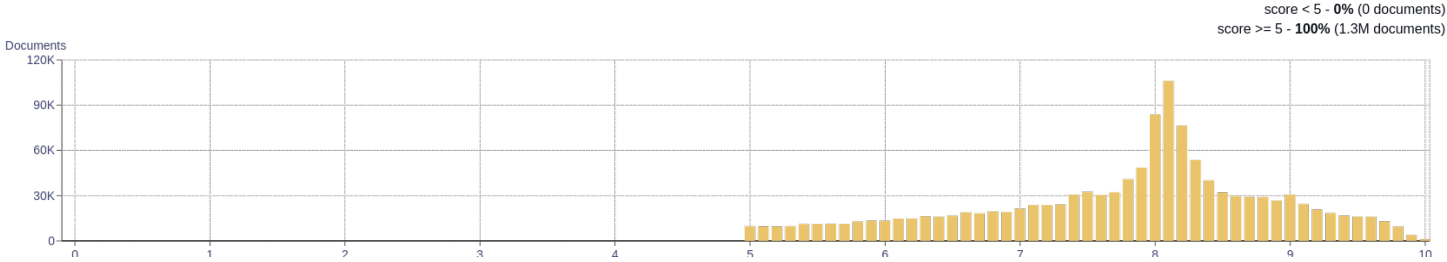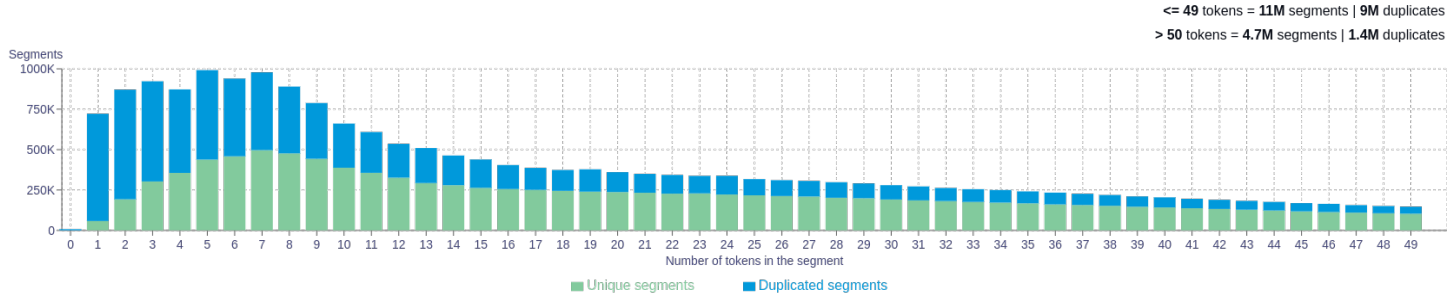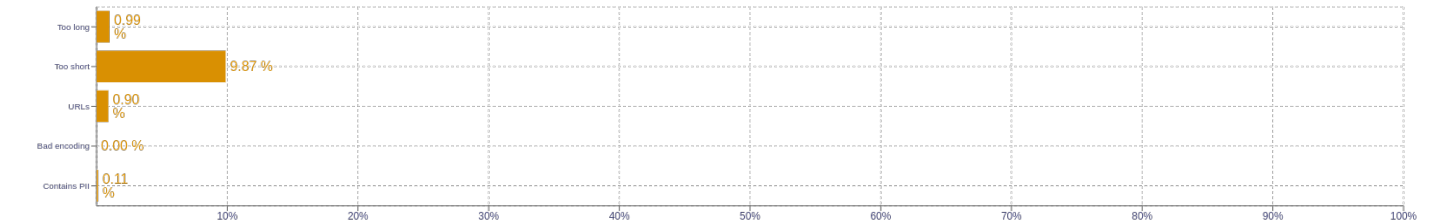
Frequent n-grams

Size	n-grams
1	билан | 3299836 бу | 2456944 бир | 2146323 учун | 2134996 тоҷикистон | 1655856
2	ҷумҳурии тоҷикистон | 661595 ўзбекистон республикаси | 603614 эмомалӣ раҳмон | 170101 шаҳри душанбе | 159888 соллаллоҳу алайҳи | 147219
3	президенти ҷумҳурии тоҷикистон | 124169 вайп вайп вайп | 82986 муҳтарам эмомалӣ раҳмон | 75996 ўзбекистон республикаси олий | 74878 ҳукумати ҷумҳурии тоҷикистон | 74181
4	вайп вайп вайп вайп | 82660 асосгузори сулҳу вахдати миллӣ | 46294 ўзбекистон республикаси вазирлар маҳкамасининг | 46155 ўзбекистон 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