HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
kir_Cyrl.jsonl.tsv	9/22/2024	Kyrgyz (ky)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
676,111	10,041,028	6,263,608 (62.38 %)	312M	3.27 GB	1,916,124,485

Top 10 domains

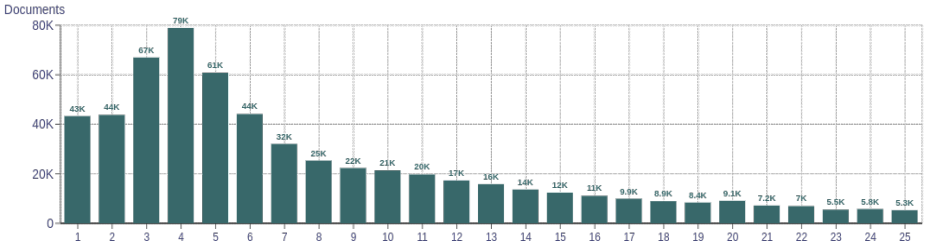
Domain	Docs	% of total
azattyk.org	201K	29.69
wikipedia.org	54K	7.92
turmush.kg	22K	3.19
sputnik.kg	12K	1.71
kyrgyztoday.org	10K	1.48
kabar.kg	9K	1.34
tyup.net	7.2K	1.07
bagyt.kg	6.6K	0.98
ykt.ru	6.6K	0.97
centralasian.org	6.2K	0.91

Top 10 TLDs

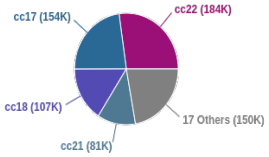
Domain	Docs	% of total
org	293K	43.29
kg	215K	31.77
com	54K	7.93
ru	45K	6.59
net	9.9K	1.46
gov.kg	8.6K	1.27
media	6.2K	0.92
asia	5.4K	0.79
info	5.3K	0.78
news	4.1K	0.61

Documents size (in segments)

<= 25 segments **88.18%** (596K documents)
> 25 segments **11.82%** (80K documents)

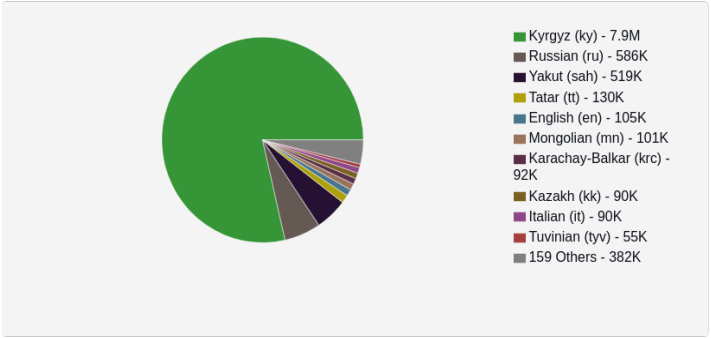


Documents by collection

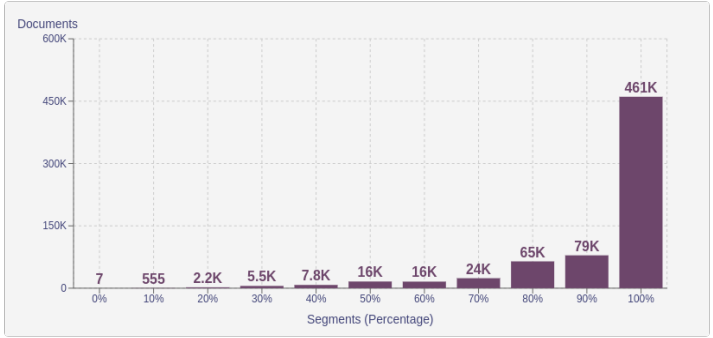


Language Distribution

Number of segments

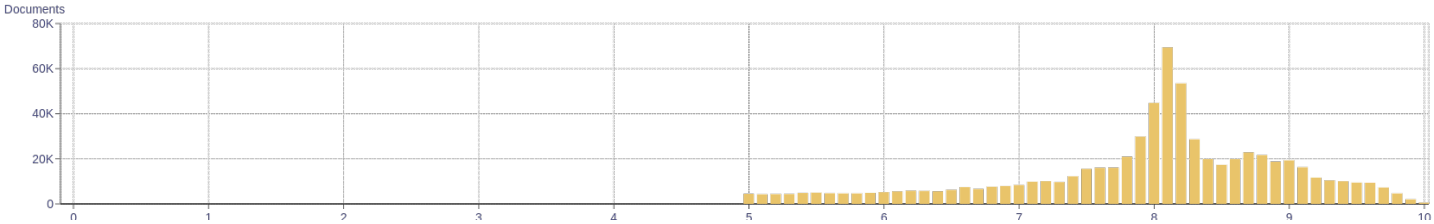


Percentage of segments in Kyrgyz (ky) inside documents



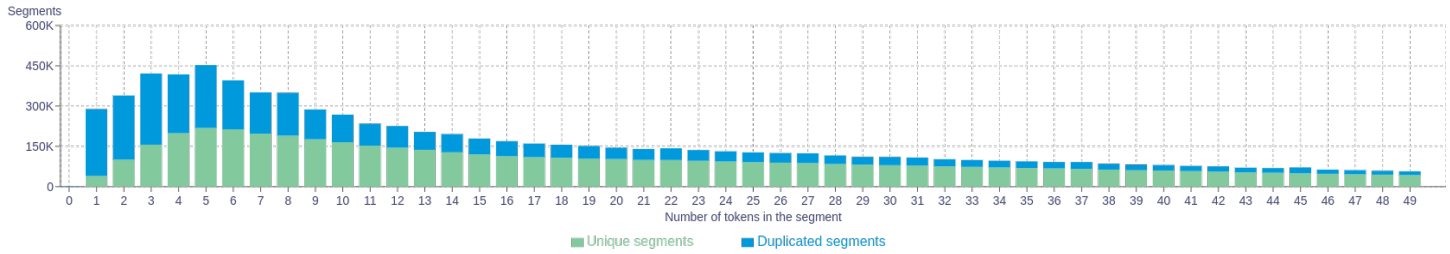
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (676K documents)

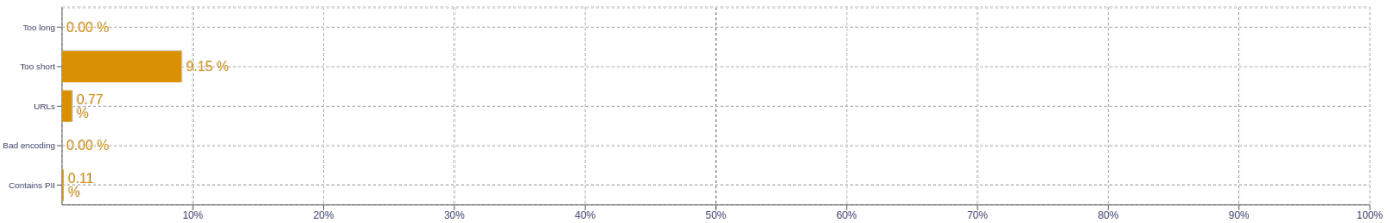


Segment length distribution by token

<= 49 tokens = **4.9M** segments | **3.3M** duplicates
> 50 tokens = **1.8M** segments | **496K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	жылы | 356235 · иш | 344897 · а | 340362 · республикасынын | 333084 · жөнүндө | 329113
2	билим берүү | 90359 · орун басары | 63181 · булагын өңдоо | 53054 · жылдан бери | 52625 · ички иштер | 48316
3	жергиликтүү өз алдынча | 23652 · тил жана энциклопедия | 19418 · республикасынын жогорку кеңешинин | 17184 · тышкы иштер министри | 15110 · президент алмазбек атамбаев | 12351
4	тил жана энциклопедия борбору | 19395 · жергиликтүү өз алдынча башкаруу | 19337 · билим берүү жана илим | 11521 · тик так тик так | 8035 · так тик так тик | 8035
5	тик так тик так тик | 8030 · так тик так тик так | 8020 · жергиликтүү өз алдынча башкаруу органдарынын | 5659 · жергиликтүү өз алдынча башкаруу органдары | 4119 · билим берүү жана илим министрлигинин | 3473

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