HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
ky_1.jsonl.tsv	3/17/2024	Kyrgyz (ky)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
88,322	11,748,340	2,695,628 (22.94 %)	132M	1.29 GB	

Top 10 domains

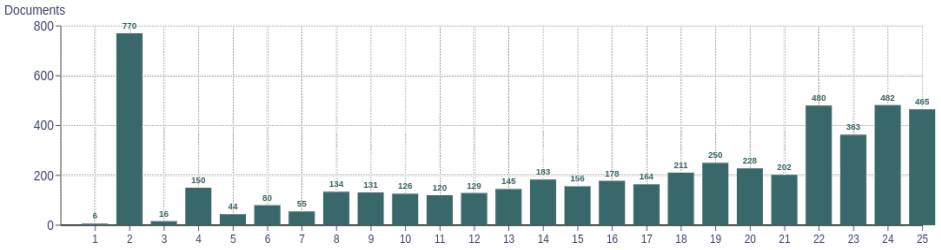
Domain	Docs	% of total
presskg.com	27K	30.88
kyrgyztoday.org	5.8K	6.60
azattyk.org	2.9K	3.33
wikipedia.org	2.9K	3.24
turmush.kg	2.1K	2.40
birge.info	1.8K	2.00
nazarnews.kg	1.7K	1.95
bagyt.kg	1.3K	1.46
tyup.net	1.2K	1.33
reporter.kg	1.1K	1.25

Top 10 TLDs

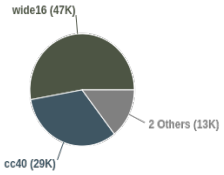
Domain	Docs	% of total
com	32K	35.76
kg	25K	28.63
org	14K	16.41
ru	2.8K	3.19
info	2.4K	2.74
news	1.9K	2.20
net	1.7K	1.98
gov.kg	1.6K	1.83
biz	1.1K	1.26
asia	1.1K	1.20

Documents size (in segments)

<= 25 segments **5.97%** (5.3K documents)
> 25 segments **94.03%** (83K documents)

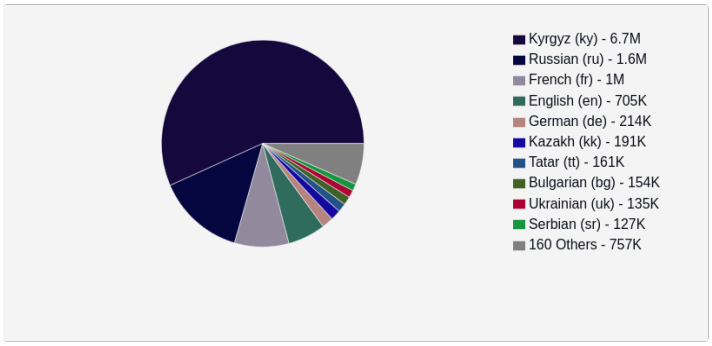


Documents by collection

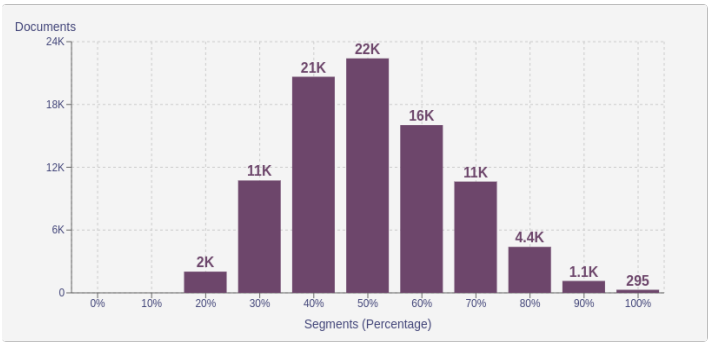


Language Distribution

Number of segments

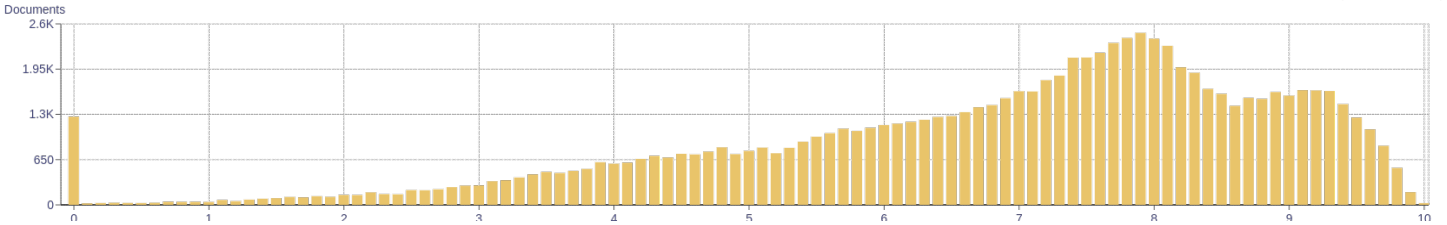


Percentage of segments in Kyrgyz (ky) inside documents



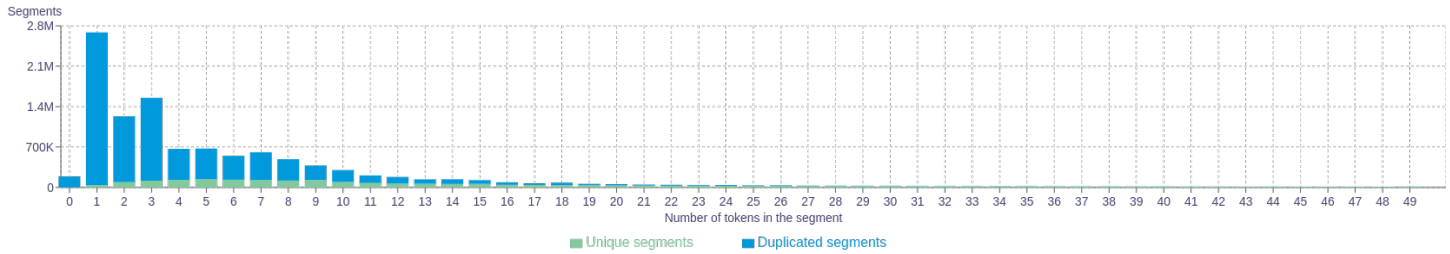
Distribution of documents by document score

score < 5 - **17.83%** (16K documents)
score >= 5 - **82.17%** (73K documents)

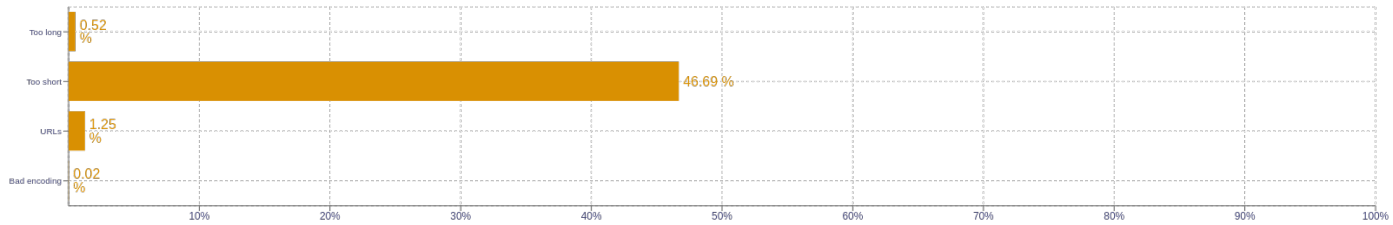


Segment length distribution by token

<= 49 tokens = **2.2M** segments | **9M** duplicates
> 50 tokens = **550K** segments | **65K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	бет | 382383 · kg | 234760 · кыргызстан | 182462 · жөнүндө | 168895 · республикасынын | 150480
2	билим берүү | 33791 · кат келиптир | 25909 · курманбек бакиев | 25701 · аскар акаев | 24623 · министрликтин жообу | 23172
3	министрликтин жаш кадрларынын | 23162 · жаш кадрларынын жоруктары | 23162 · үнү тарыхый мурас | 13710 · эркинтоо биримдик пресс | 13710 · энесай новости иссык | 13710
4	министрликтин жаш кадрларынын жоруктары | 23162 · өкмөтү нур эл de | 13710 · үнү тарыхый мурас обон | 13710 · эркинтоо биримдик пресс kg | 13710 · экспресс саратан diezel айыл | 13710
5	өкмөтү нур эл de факто | 13710 · үнү тарыхый мурас обон ош | 13710 · эркинтоо биримдик пресс kg нур | 13710 · экспресс саратан diezel айыл өкмөтү | 13710 · шамы ош жаңырыгы кыргыз руху | 13710

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