HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
crh_Latn.jsonl.tsv	11/27/2024	Crimean Tatar (crh)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
122,744	1,380,903	774,648 (56.10 %)	46M	301.82 MB	279,816,188

Top 10 domains

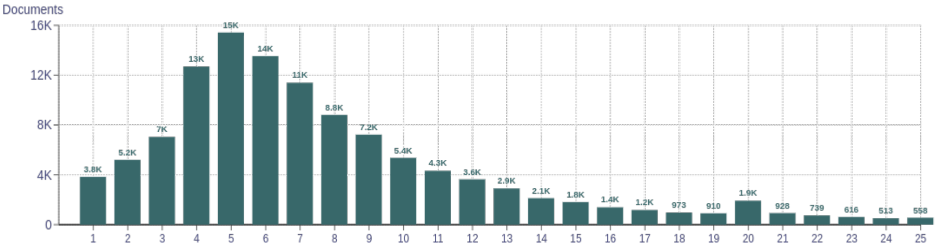
Domain	Docs	% of total
krymr.com	53K	43.04
azatliq.org	22K	17.80
wikipedia.org	14K	11.63
inform.kz	4.9K	3.97
trt.net.tr	2.4K	1.95
qazaqtimes.com	2.3K	1.88
almaty-akshamy.kz	2K	1.59
avdet.org	1.5K	1.23
turkuindir.info	1.4K	1.14
baq.kz	996	0.81

Top 10 TLDs

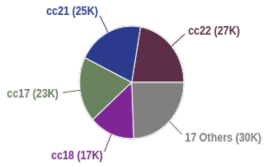
Domain	Docs	% of total
com	60K	48.62
org	39K	32.00
kz	16K	12.69
net.tr	2.4K	1.95
info	1.5K	1.21
net	890	0.73
ru	626	0.51
uz	613	0.50
gov.ua	524	0.43
az	295	0.24

Documents size (in segments)

<= 25 segments **93.72%** (115K documents)
> 25 segments **6.28%** (7.7K documents)

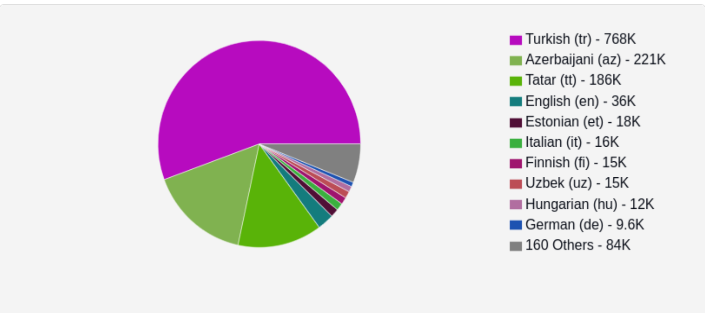


Documents by collection



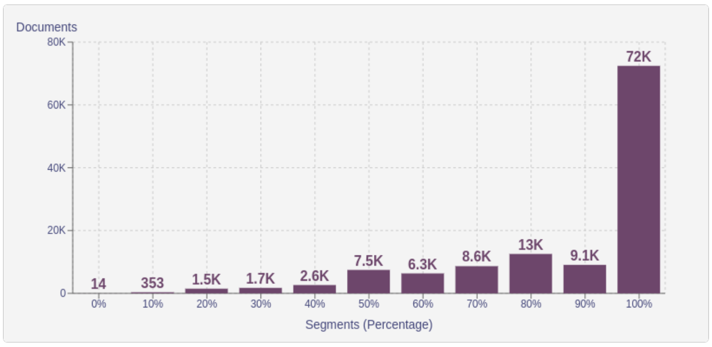
Language Distribution

Number of segments



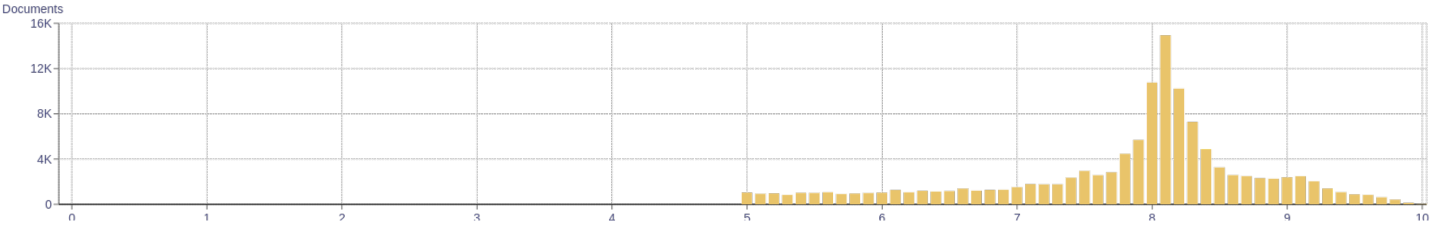
*Crimean Tatar (crh) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Crimean Tatar (crh) inside documents



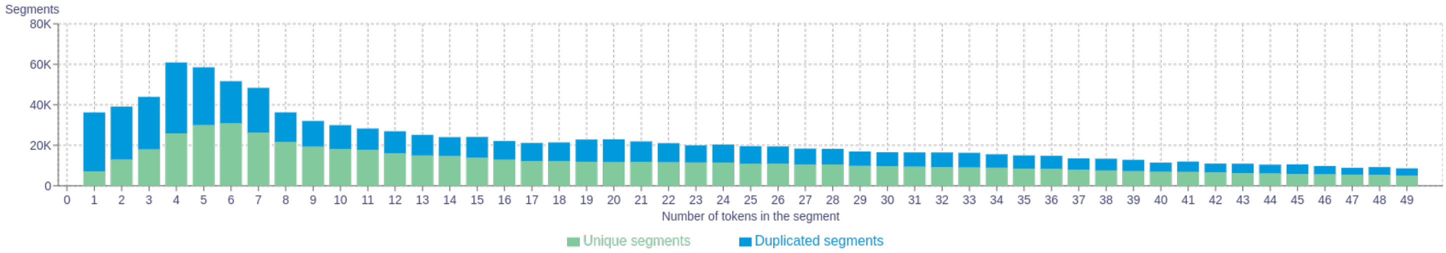
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (123K documents)

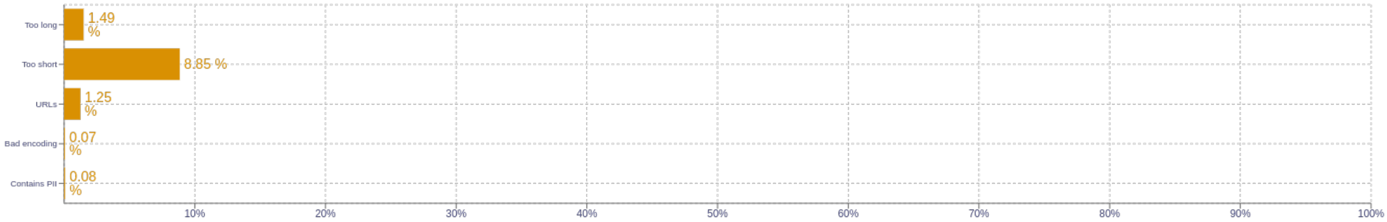


Segment length distribution by token

<= 49 tokens = **593K** segments | **511K** duplicates
> 50 tokens = **276K** segments | **95K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	qırım | 157006 rusiye | 147118 dep | 144398 häm | 121927 edi | 111291
2	qırımtatar milliy | 21231 hizb ut | 18252 işğal etilgen | 17655 ekrem çelebi | 15204 qayd etti | 12629
3	qırımtatar milliy meclisiniñ | 7628 ukraina prezidenti petro | 6895 prezidenti petro poroşenko | 6807 tarihiy adaletniñ tiklenmesi | 6422 sıra iqtisadiy sanktsiyalarnı | 5954
4	ukraina prezidenti petro poroşenko | 6657 ğarp memleketleri bir sıra | 4693 memleketleri bir sıra iqtisadiy | 4693 rusiye prezidenti vladimir putin | 4500 prezidenti petro poroşenko bunıñnen | 4390
5	ğarp memleketleri bir sıra iqtisadiy | 4693 ukraina prezidenti petro poroşenko bunıñnen | 4390 prezidenti petro poroşenko bunıñnen bağlı | 4390 poroşenko bunıñnen bağlı qanunnı imzaladı | 4385 petro poroşenko bunıñnen bağlı qanunnı | 4382

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