HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
tat_Cyrl.jsonl.tsv	9/16/2024	Tatar (tt)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
630,685	13,448,632	6,363,903 (47.32 %)	381M	3.63 GB	2,143,760,119

Top 10 domains

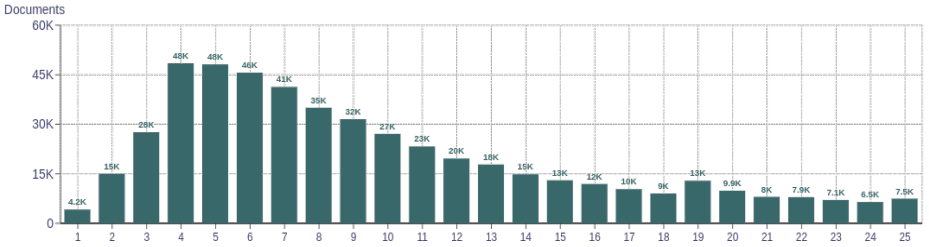
Domain	Docs	% of total
azatliq.org	66K	10.48
wikipedia.org	62K	9.88
tatar-inform.tatar	16K	2.48
shahrikazan.ru	13K	2.08
syuyumbike.ru	13K	2.02
matbugat.ru	10K	1.59
tatar-congress.org	9.2K	1.47
kazanutlary.ru	8.4K	1.33
arskmedia.ru	8K	1.28
alabuganury.ru	8K	1.27

Top 10 TLDs

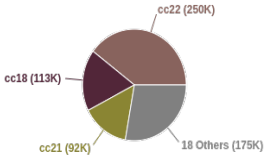
Domain	Docs	% of total
ru	407K	64.51
org	148K	23.43
com	31K	4.99
tatar	25K	3.92
info	3.4K	0.54
xn--p1ai	3.1K	0.49
su	2.6K	0.41
net	2K	0.32
net.tr	1.7K	0.27
co	700	0.11

Documents size (in segments)

<= 25 segments **79.83%** (503K documents)
> 25 segments **20.17%** (127K documents)

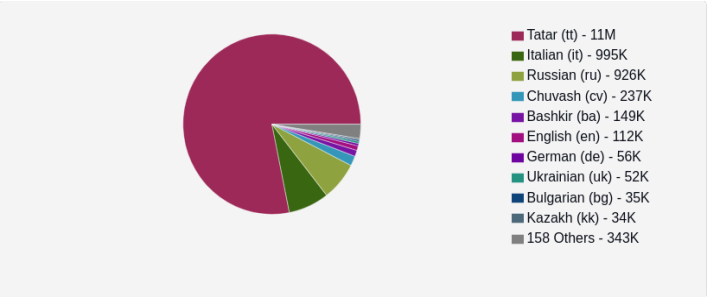


Documents by collection

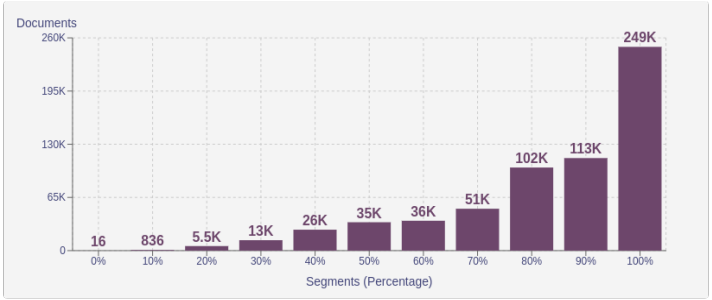


Language Distribution

Number of segments

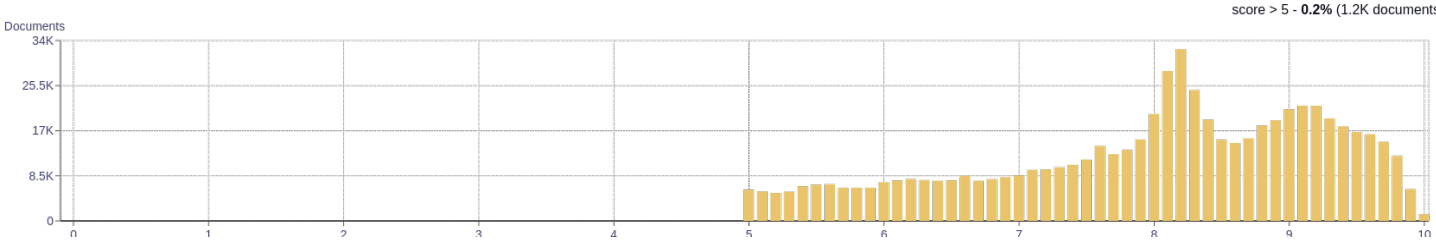


Percentage of segments in Tatar (tt) inside documents



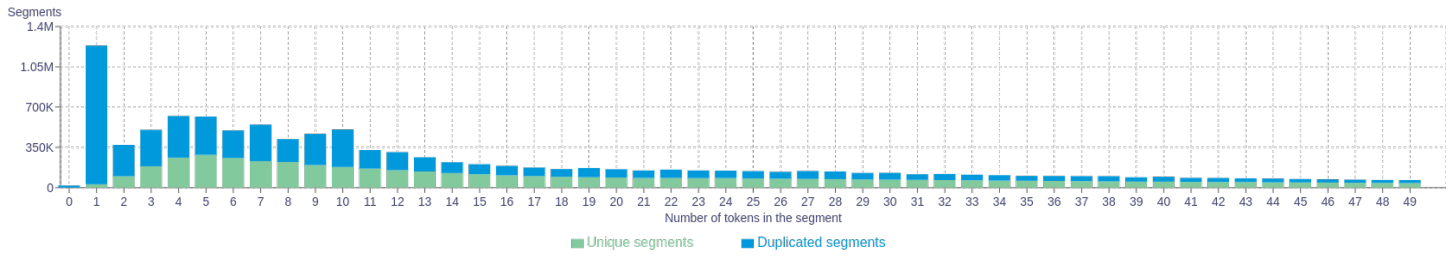
Distribution of documents by document score

score <= 5 - **99.8%** (629K documents)
score > 5 - **0.2%** (1.2K documents)

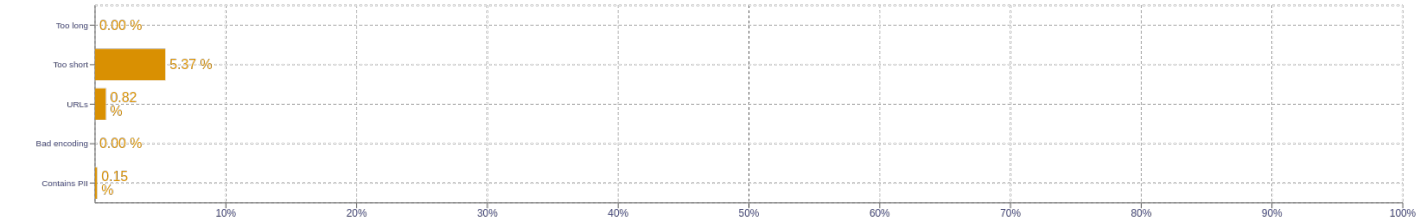


Segment length distribution by token

<= 49 tokens = **4.9M** segments | **6.2M** duplicates
> 50 tokens = **2.3M** segments | **865K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	татар | 890689 · булган | 557195 · кеше | 524411 · зур | 500782 · алып | 490714
2	текстны үзгәртү | 150989 · татарстан республикасы | 97054 · авыл хуҗалыгы | 84316 · в telegram | 82214 · следите за | 82191
3	самым важным и | 82177 · следите за самым | 82176 · интересным в telegram | 82176 · и интересным в | 82176 · за самым важным | 82176
4	следите за самым важным | 82176 · самым важным и интересным | 82176 · и интересным в telegram | 82176 · за самым важным и | 82176 · важным и интересным в | 82176
5	следите за самым важным и | 82176 · самым важным и интересным в | 82176 · за самым важным и интересным | 82176 · важным и интересным в telegram | 82176 · məĥim həm kызыклы язмаларны татмедиа | 55426

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