HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
tt_1.jsonl.tsv	3/17/2024	Tatar (tt)

Volumes

Docs	Segments	Unique segments	Tokens	Size
65,152	8,571,604	18,193 (0.21 %)	102M	909.93 MB

Type-Token Ratio

Tatar (tt)
0.02

Top 10 domains

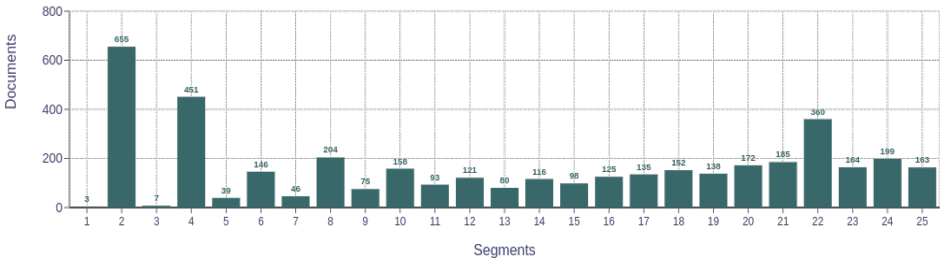
Domain	Docs	% of total
matbugat.ru	7.1K	10.91
wikipedia.org	4.7K	7.26
kiziltan.ru	2.6K	4.02
syuyumbike.ru	2.4K	3.75
tatar-today.ru	2.2K	3.32
belem.ru	2K	3.05
azatliq.org	1.7K	2.66
erlar.ru	1.5K	2.36
tatar-inform.tatar	1.5K	2.32
madanizhomga.ru	1K	1.60

Top 10 TLDs

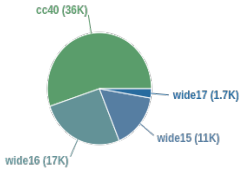
Domain	Docs	% of total
ru	49K	75.45
org	7.8K	12.04
tatar	3.3K	5.10
com	2.9K	4.50
xn--p1ai	378	0.58
su	265	0.41
show	230	0.35
net	145	0.22
info	103	0.16
biz	93	0.14

Documents size (in segments)

<= 25 segments **6.27%** (4.1K documents)
> 25 segments **93.73%** (61K documents)

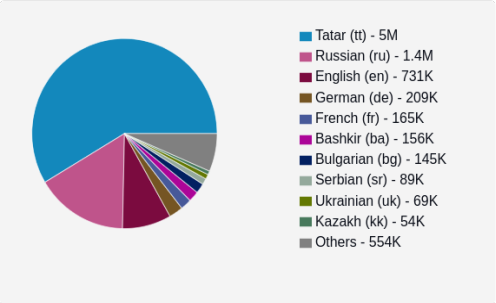


Documents by collection

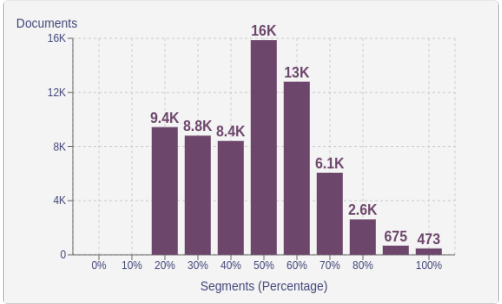


Language Distribution

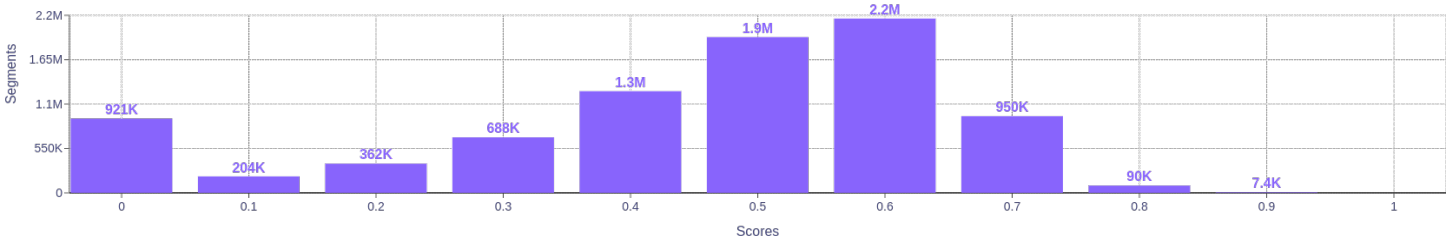
Number of segments



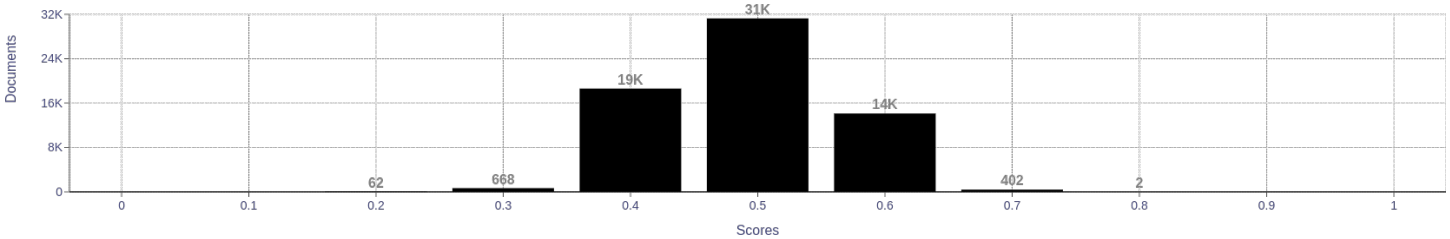
Percentage of segments in Tatar (tt) inside documents



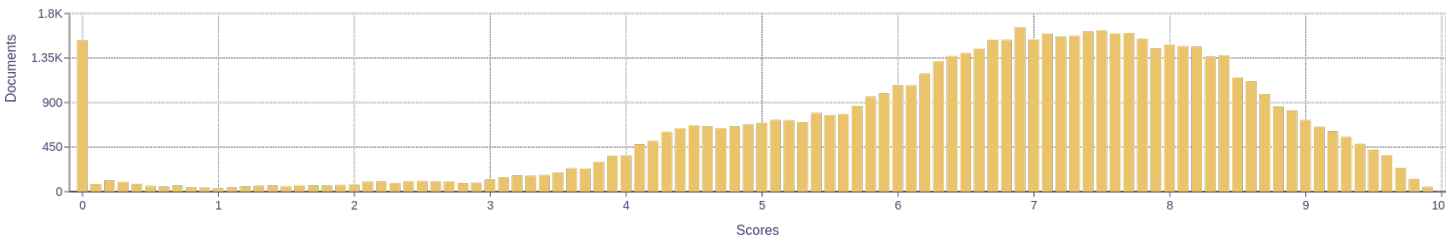
Distribution of segments by fluency score



Distribution of documents by average fluency score



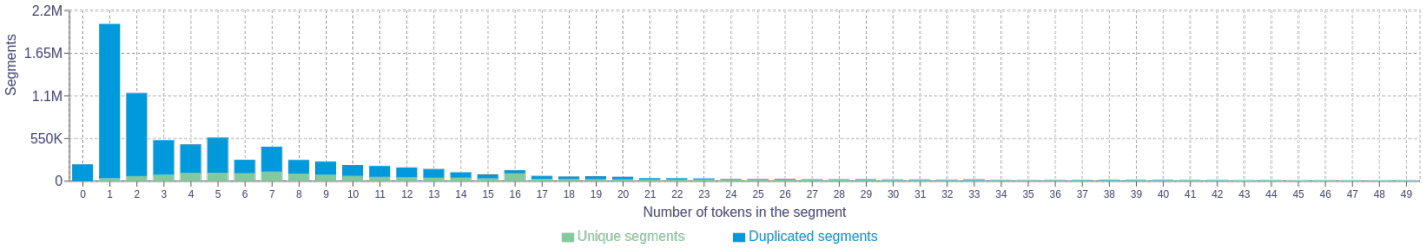
Distribution of documents by document score



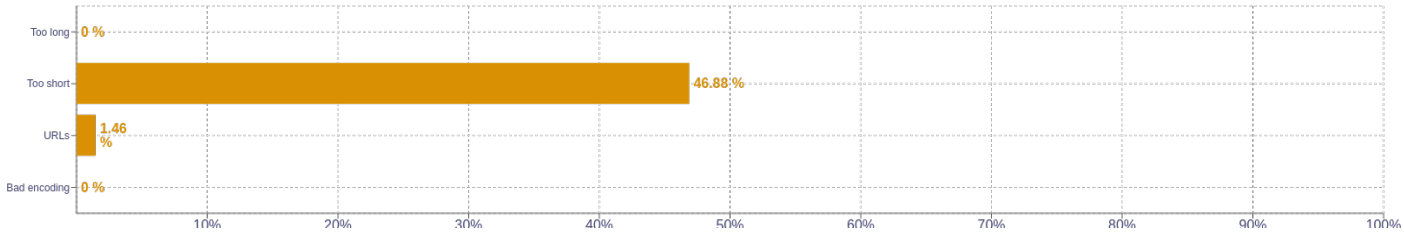
Segment length distribution by token

<= 49 tokens = **1.7M** segments | **6.4M** duplicates

> 50 tokens = **449K** segments | **79K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	татар | 311276 · в | 215765 · и | 207035 · татарстан | 140509 · яңа | 139566
2	на сайте | 55449 · татарстан республикасы | 25753 · татар теле | 23243 · авыл хуҗалыгы | 21971 · все права | 19957
3	все права защищены | 19854 · размещенные на сайте | 19382 · только с письменного | 19369 · с письменного согласия | 19361 · размещенной на сайте | 19268
4	только с письменного согласия | 19360 · в любом объеме информации | 19265 · с письменного согласия редакций | 19250 · письменного согласия редакций сми | 19250 · возможна только с письменного | 19250
5	только с письменного согласия редакций | 19250 · с письменного согласия редакций сми | 19250 · возможна только с письменного согласия | 19250 · распространение в любом объеме информации | 19229 · и распространение в любом объеме | 19229

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