HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
bak_Cyrl.jsonl.tsv	9/24/2024	Bashkir (ba)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
170,822	3,138,502	1,714,242 (54.62 %)	94M	957.58 MB	555,532,259

Top 10 domains

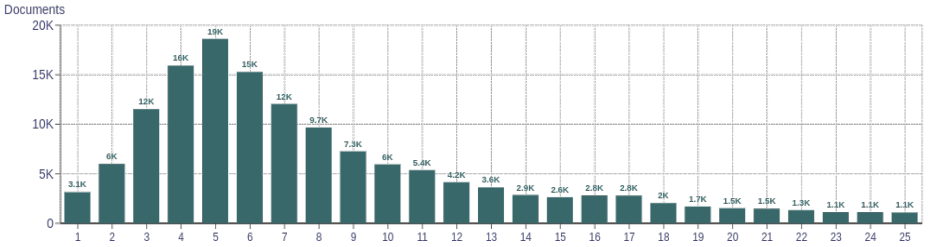
Domain	Docs	% of total
bashinform.ru	36K	21.23
wikipedia.org	34K	19.72
hakmar.ru	6.6K	3.84
bashgazet.ru	6.1K	3.57
башкирская-энциклопедия.рф	5.4K	3.13
ural-rb.ru	3.4K	2.01
bashnews.eu	3.1K	1.82
ye02.ru	3.1K	1.79
tv-rb.ru	2.3K	1.37
phones-expo.news	2K	1.20

Top 10 TLDs

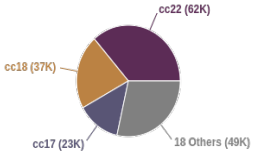
Domain	Docs	% of total
ru	103K	60.07
org	37K	21.41
com	11K	6.43
рф	6.6K	3.88
info	4.1K	2.38
eu	3.2K	1.86
news	2.3K	1.36
su	1.3K	0.74
tv	744	0.44
net	423	0.25

Documents size (in segments)

<= 25 segments **82.62%** (141K documents)
> 25 segments **17.38%** (30K documents)

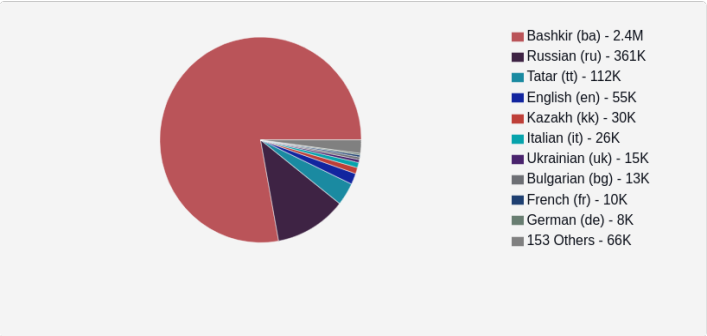


Documents by collection

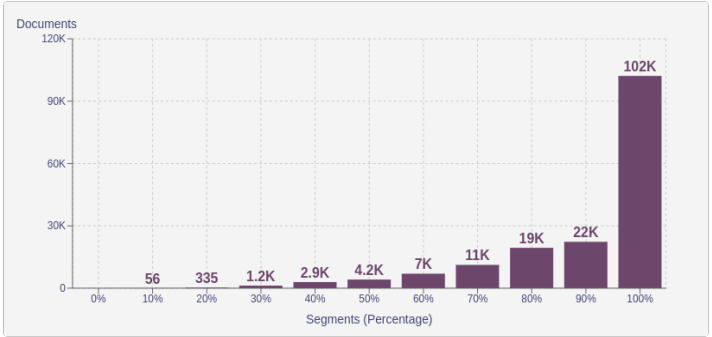


Language Distribution

Number of segments

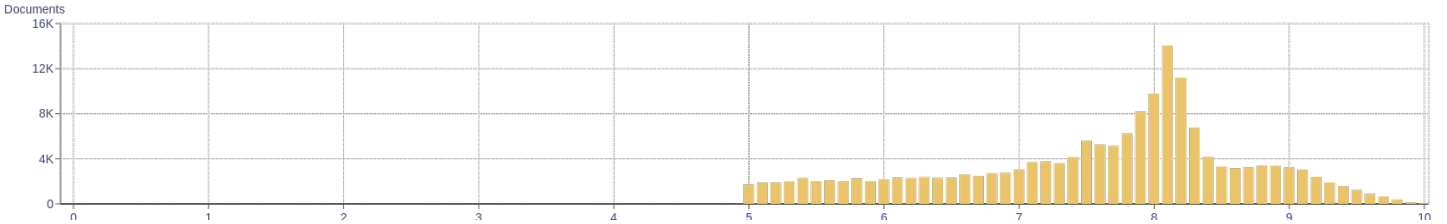


Percentage of segments in Bashkir (ba) inside documents



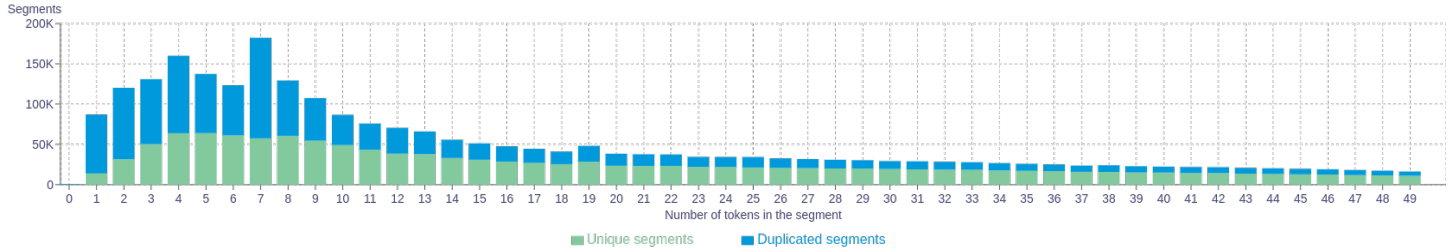
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (171K documents)

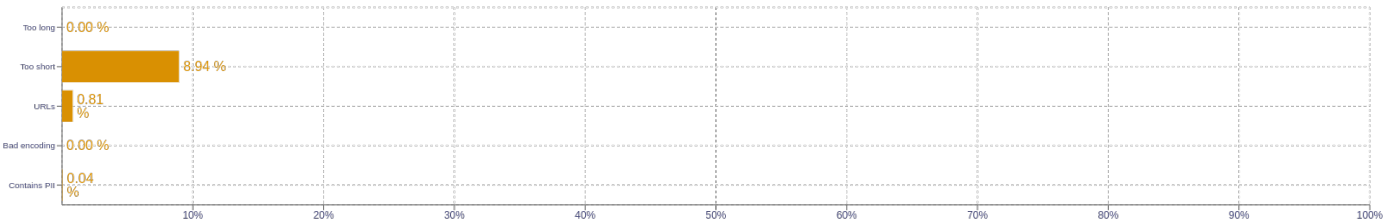


Segment length distribution by token

<= 49 tokens = **1.3M** segments | **1.2M** duplicates
> 50 tokens = **570K** segments | **176K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	бер | 394643 · ҙҙгәртергә | 333478 · йылда | 168648 · ук | 166049 · башҡорт | 156499
2	вики-тексты ҙҙгәртергә | 144918 · бер нисә | 40357 · ауыл хуҙалығы | 29362 · ҡайһы бер | 28608 · башҡортостан республикаһының | 25604
3	сығанаҡ кодты ҙҙгәртергә | 9083 · төйәк ватан һуғышында | 7949 · төйәк ватан һуғышы | 7460 · тәүге сығанаҡтан архивланған | 7223 · башҡортостан республикаһының атҡаҙанған | 7161
4	башҡортостан президенты рәстәм хәмитов | 3634 · һанына тамамланған йылдарҙа тыуғандар | 3617 · башҡорт теле һәм әҙәбиәте | 3408 · брокгауз һәм ефрондың энциклопедик | 3014 · сығышы менән хәҙерге башҡортостан | 2886
5	брокгауз һәм ефрондың энциклопедик һүҙлеге | 2987 · сығышы менән хәҙерге башҡортостан республикаһының | 2711 · сайттарында тура актив гиперһылтанма ҡуйырға | 2659 · интернет сайттарында тура актив гиперһылтанма | 2659 · хеҙмәт һәм халыҡты социаль яҡлау | 2070

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