

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

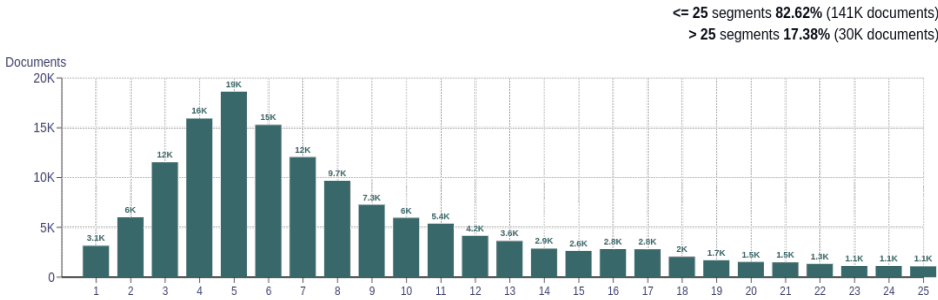
Top 10 domains

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

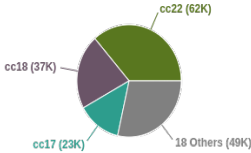
Top 10 TLDs

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)

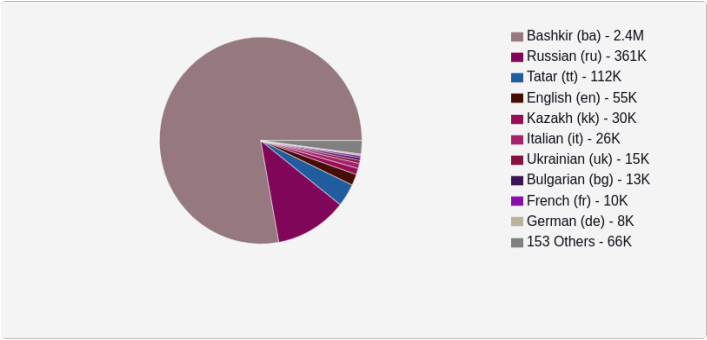


Documents by collection

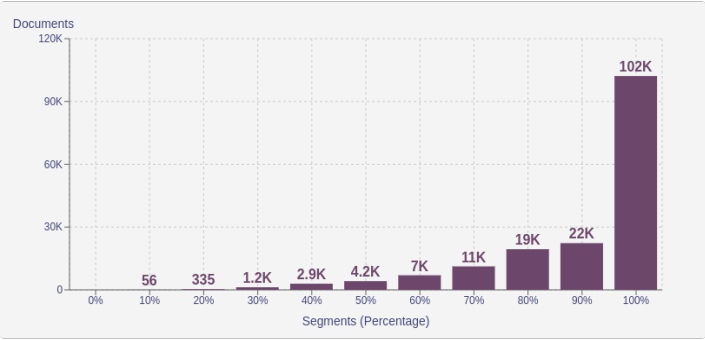


Language Distribution

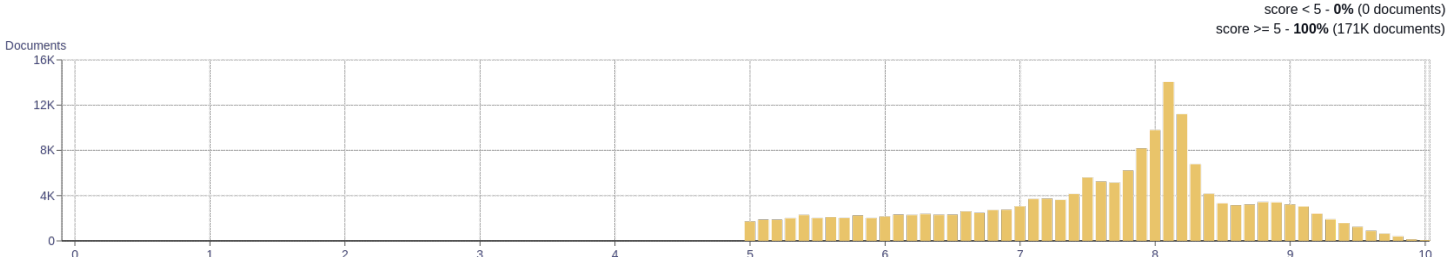
Number of segments



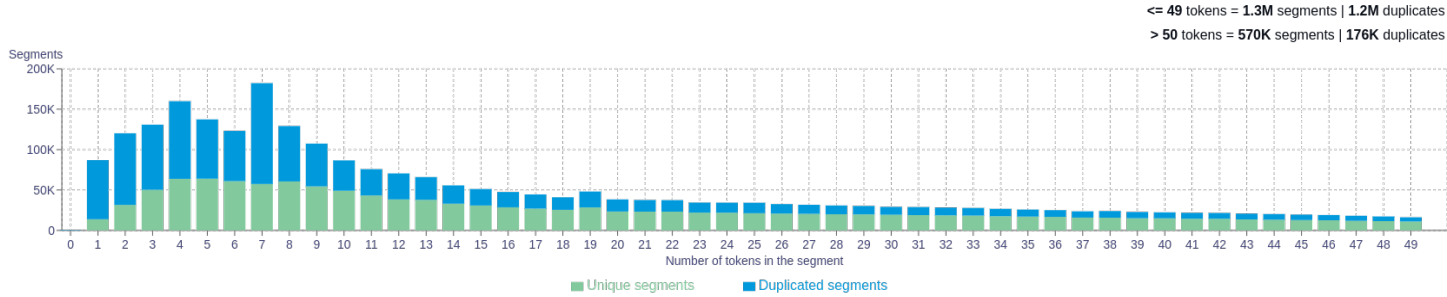
Percentage of segments in Bashkir (ba) inside documents



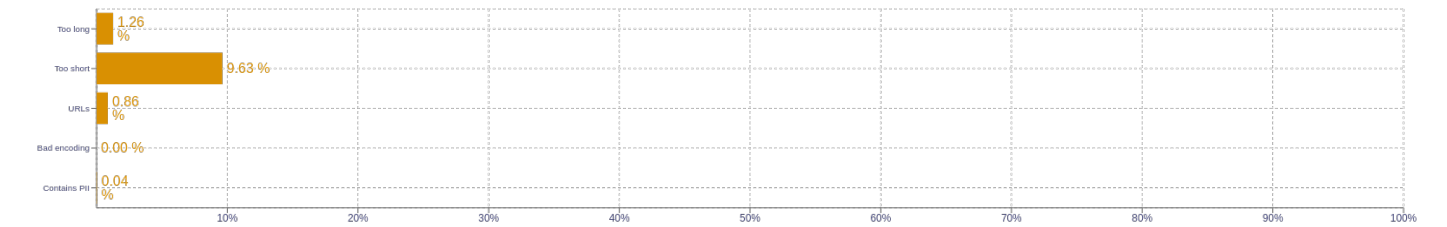
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>бер 394643</div> <div>үзгәртәргә 333478</div> <div>йылда 168648</div> <div>ук 166049</div> <div>башкорт 156499</div>
2	<div>вики-тексты үзгәртәргә 144918</div> <div>бер нисә 40357</div> <div>ауыл хужалыгы 29362</div> <div>хайһы бер 28608</div> <div>башкортостан республикаһының 25604</div>
3	<div>сығанаҡ кодты үзгәртәргә 9083</div> <div>бөйөк ватан һуғышында 7949</div> <div>бөйөк ватан һуғышы 7460</div> <div>тәүге сығанаҡтан архивланған 7223</div> <div>башкортостан республикаһының атказанған 7161</div>
4	<div>башкортостан президенты рөстәм хәмитов 3634</div> <div>һанына тамамланған йылдарҙа тыуғандар 3617</div> <div>башкорт теле һәм әҙәбиәте 3408</div> <div>брокгауз һәм ефрондың энциклопедик 3014</div> <div>сығышы менән хәҙерге башкортостан 2886</div>
5	<div>брокгауз һәм ефрондың энциклопедик һүзлегә 2987</div> <div>сығышы менән хәҙерге башкортостан республикаһының 2711</div> <div>сайттарында тура актив гиперһылтанма хуйырға 2659</div> <div>интернет сайттарында тура актив гиперһылтанма 2659</div> <div>хезмәт һәм халыҡты социаль яҡлау 2070</div>

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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.stinfo.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>, , , 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>