HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
be_1.jsonl.tsv	3/21/2024	Belarusian (be)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
356,534	38,016,416	10,269,652 (27.01 %)	517M	4.51 GB	

Top 10 domains

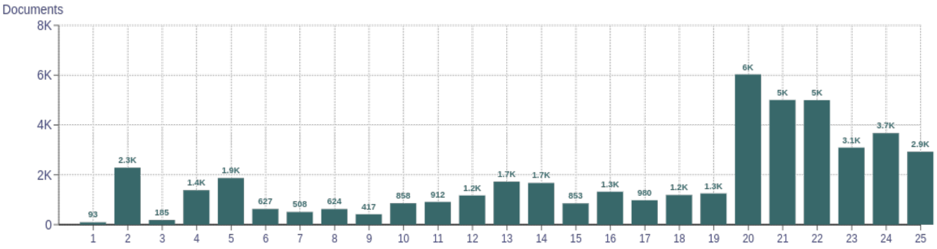
Domain	Docs	% of total
wikipedia.org	25K	6.98
skarnik.by	19K	5.20
budzma.by	10K	2.83
catholicnews.by	8.8K	2.47
slounik.org	6.4K	1.81
spring96.org	6.3K	1.77
kimpress.by	5K	1.40
googleplaystoreapks.com	3.4K	0.95
belsat.eu	3.3K	0.94
cloudfront.net	3.3K	0.93

Top 10 TLDs

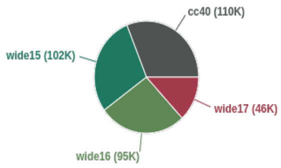
Domain	Docs	% of total
by	121K	33.83
org	67K	18.89
ru	57K	16.11
com	38K	10.62
net	14K	3.88
info	10K	2.79
eu	7.8K	2.20
in.ua	4.6K	1.30
gov.by	4.2K	1.18
fm	3.9K	1.08

Documents size (in segments)

<= 25 segments **12.93%** (46K documents)
> 25 segments **87.07%** (307K documents)

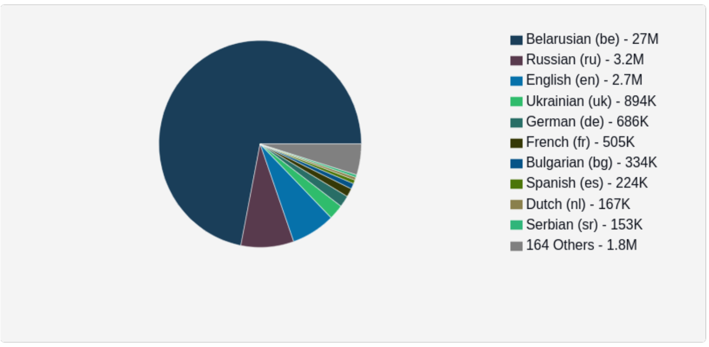


Documents by collection

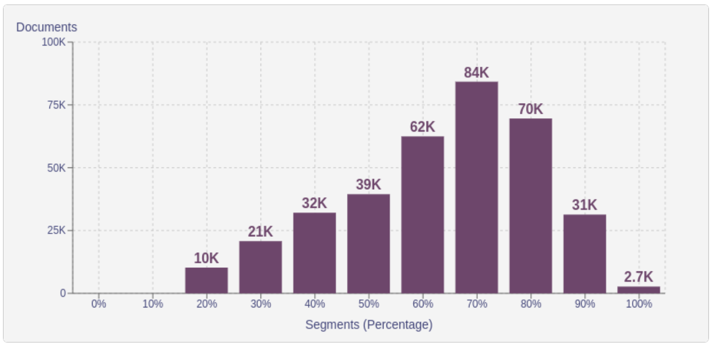


Language Distribution

Number of segments

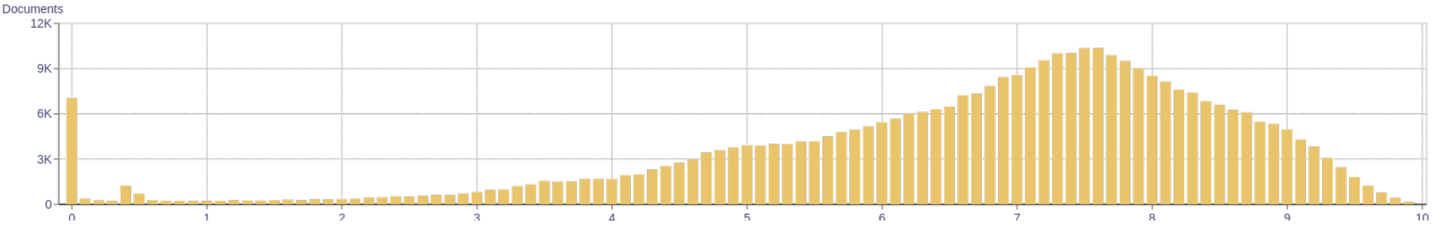


Percentage of segments in Belarusian (be) inside documents



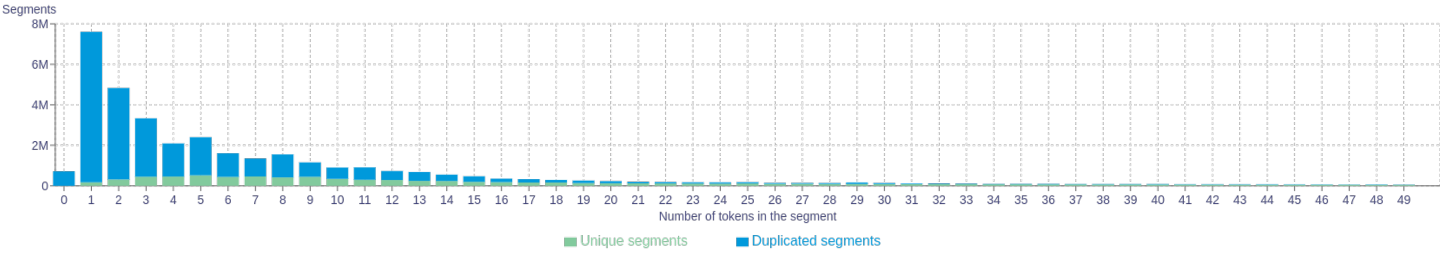
Distribution of documents by document score

score < 5 - **16.46%** (59K documents)
score >= 5 - **83.54%** (298K documents)

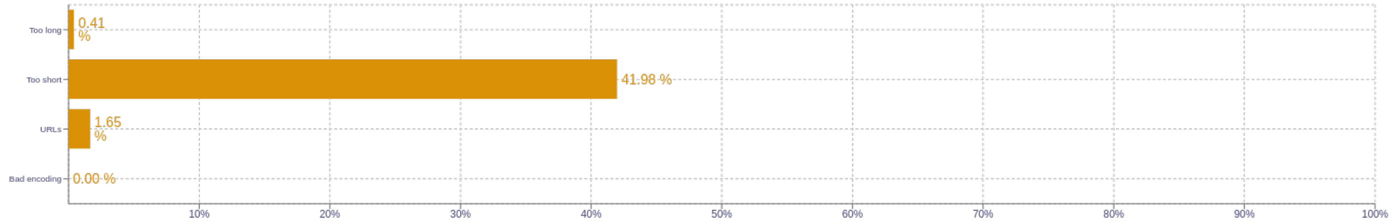


Segment length distribution by token

<= 49 tokens = **8.3M** segments | **27M** duplicates
> 50 tokens = **2.4M** segments | **436K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ў | 6176577 да | 2343371 як | 1613889 ад | 1160746 пра | 1117709
2	е ў | 150570 рэспублікі беларусь | 124645 ў беларусі | 122596 кропка расы | 115818 суадносіны тэмпературы | 105229
3	тэмпературы і вільготнасці | 105329 ападкаў не чакаецца | 70865 б в г | 64652 л м н | 63522 к л м | 63423
4	суадносіны тэмпературы і вільготнасці | 105227 к л м н | 63205 п р с т | 63107 х ц ч ш | 62529 ф х ц ч | 62292
5	ф х ц ч ш | 62215 л м н о п | 61155 к л м н о | 61112 м н о п р | 61105 н о п р с | 61009

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