

General overview

Corpus	Analytics date	Language
bel_Cyrl.tsv	9/16/2024	Belarusian (be)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
2,319,619	48,844,196	23,307,246 (47.72 %)	1.5B	14.23 GB	8,493,165,091

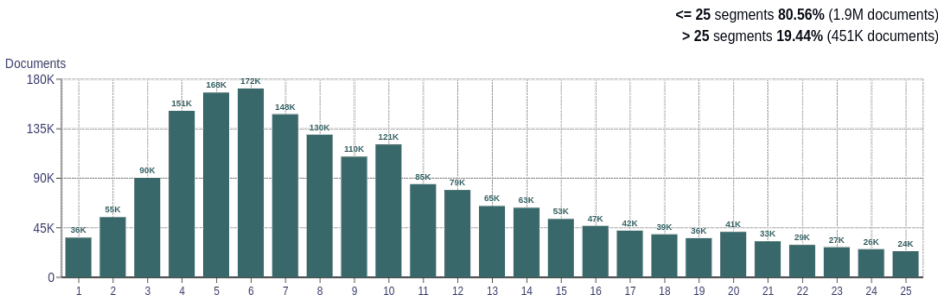
Top 10 domains

Domain	Docs	% of total
wikipedia.org	407K	17.54
svaboda.org	219K	9.43
spring96.org	87K	3.73
racyja.com	62K	2.66
nashaniva.com	62K	2.66
belsat.eu	40K	1.73
novychas.by	34K	1.46
zviazda.by	27K	1.15
skarnik.by	24K	1.02
nn.by	23K	1.01

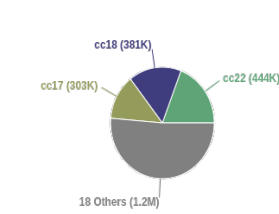
Top 10 TLDs

Domain	Docs	% of total
org	858K	36.98
by	632K	27.24
com	337K	14.55
ru	126K	5.44
net	72K	3.09
eu	63K	2.74
info	55K	2.39
gov.by	32K	1.38
fm	25K	1.06
online	16K	0.71

Documents size (in segments)

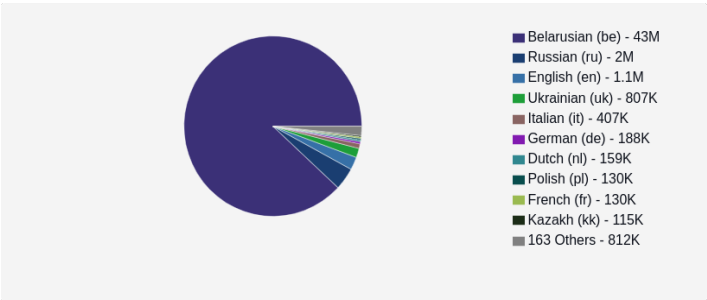


Documents by collection

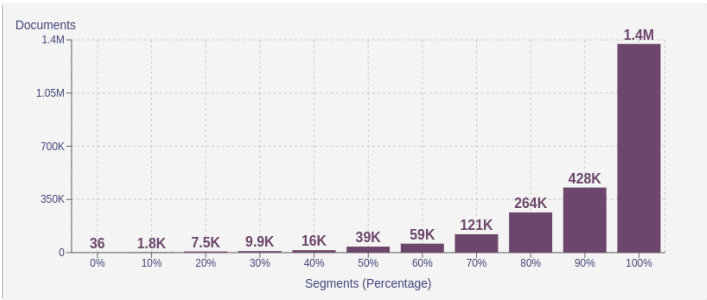


Language Distribution

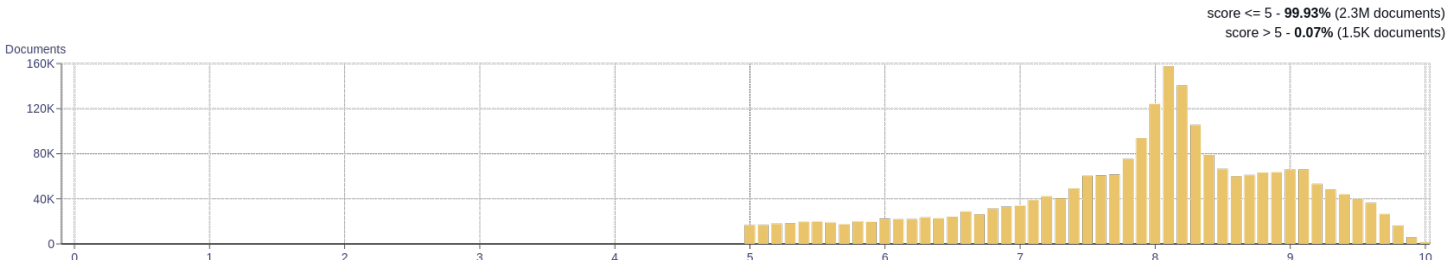
Number of segments



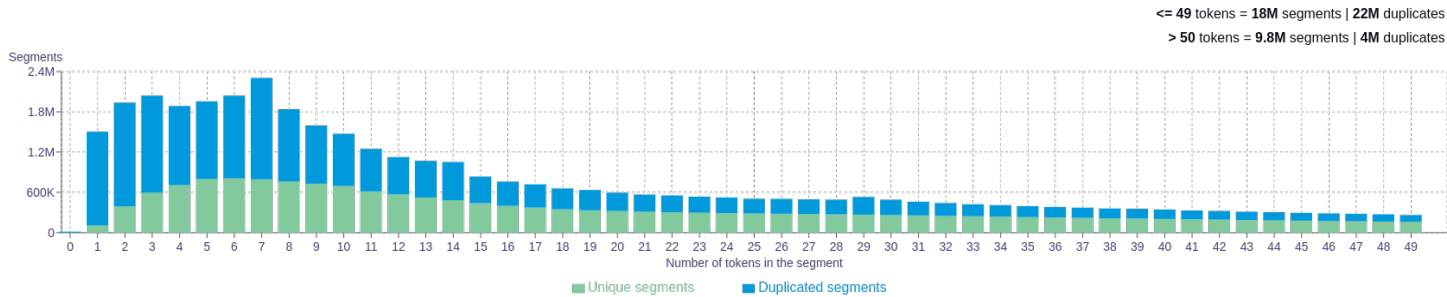
Percentage of segments in Belarusian (be) inside documents



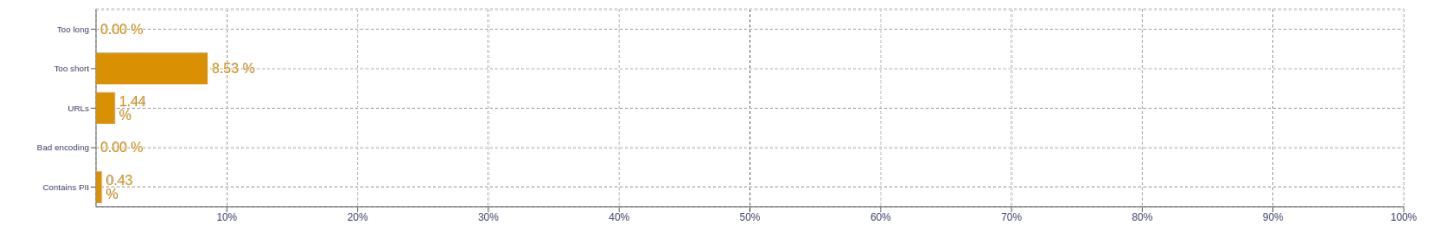
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>ў   22444103<div>да   7444290</div>як   5589025<div>гэта   4896701</div>ад   4025399</div>
2	<div>правіць зыходнік   710669<div>рэдагаваць крыніцу   396147</div>ў беларусі   388317<div>рэспублікі беларусь   379246</div>тым ліку   324698</div>
3	<div>лістапад кастрычнік верасень   126015<div>сьнежань лістапад кастрычнік   125959</div>сакавік люты студзень   123100<div>красавік сакавік люты   123093</div>травень красавік сакавік   123060</div>
4	<div>сьнежань лістапад кастрычнік верасень   125949<div>красавік сакавік люты студзень   123079</div>травень красавік сакавік люты   123047<div>чэрвень травень красавік сакавік   122839</div>ліпень чэрвень травень красавік   122628</div>
5	<div>травень красавік сакавік люты студзень   123033<div>чэрвень травень красавік сакавік люты   122826</div>ліпень чэрвень травень красавік сакавік   122601</div> <div>жнівень ліпень чэрвень травень красавік   119774<div>верасень жнівень ліпень чэрвень травень   119745</div></div>

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 of 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>