HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-v2-rus_Cyrl.tsv	10/30/2024	Russian (ru)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
884,688,865	26,292,090,560			6.37 TB	3,882,683,302,870

Top 10 domains

Domain	Docs	% of total
livejournal.com	4.8M	0.54
blogspot.com	2.5M	0.28
resheniya-sudov.ru	2.1M	0.23
blogspot.ru	2M	0.23
aif.ru	2M	0.22
wikipedia.org	1.9M	0.22
patents.su	1.8M	0.21
academic.ru	1.6M	0.18
mail.ru	1.5M	0.17
spb.ru	1.5M	0.17

Top 10 TLDs

Domain	Docs	% of total
ru	543M	61.40
com	102M	11.48
net	33M	3.68
com.ua	26M	2.94
info	25M	2.77
org	20M	2.25
ua	19M	2.16
by	15M	1.65
su	12M	1.35
kz	9.8M	1.11

Documents size (in segments)

<= 25 segments **72.96%** (645M documents)
> 25 segments **27.04%** (239M documents)
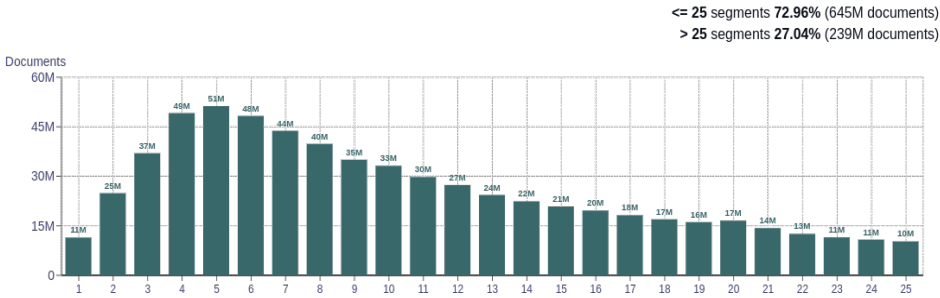
Documents by collection

cc22 (143M)
cc18 (191M)
wide15 (101M)
18 Others (449M)
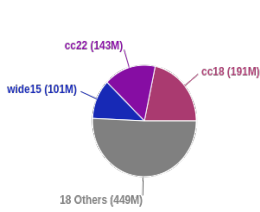
Language Distribution

Number of segments

- Russian (ru) - 24B
- English (en) - 487M
- Italian (it) - 453M
- Ukrainian (uk) - 355M
- German (de) - 130M
- French (fr) - 107M
- Serbian (sr) - 65M
- Macedonian (mk) - 37M
- Bulgarian (bg) - 34M
- Spanish (es) - 32M
- 165 Others - 343M
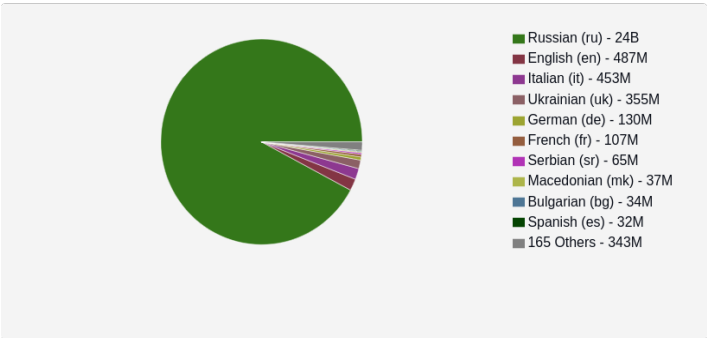
Percentage of segments in Russian (ru) inside documents
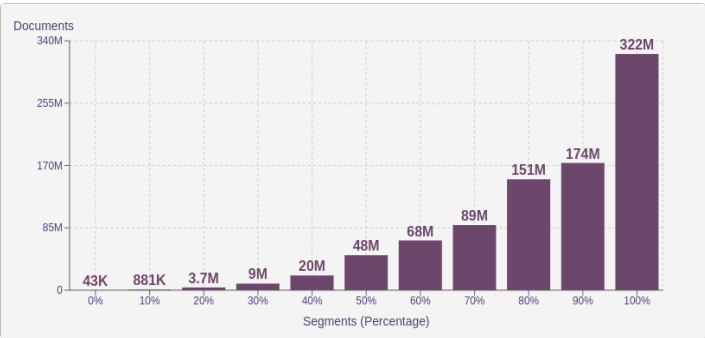
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (885M documents)

Segment noise distribution

- Too long: 0.00 %
- Too short: 7.80 %
- URLs: 1.32 %
- Bad encoding: 0.00 %
- Contains PII: 0.17 %
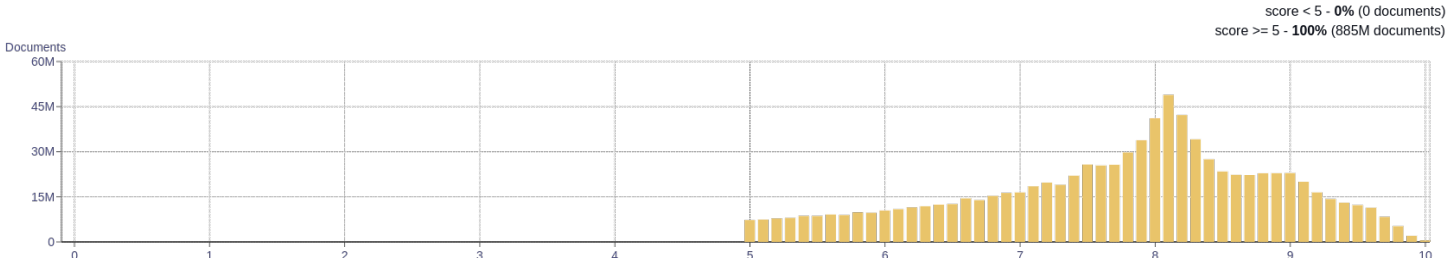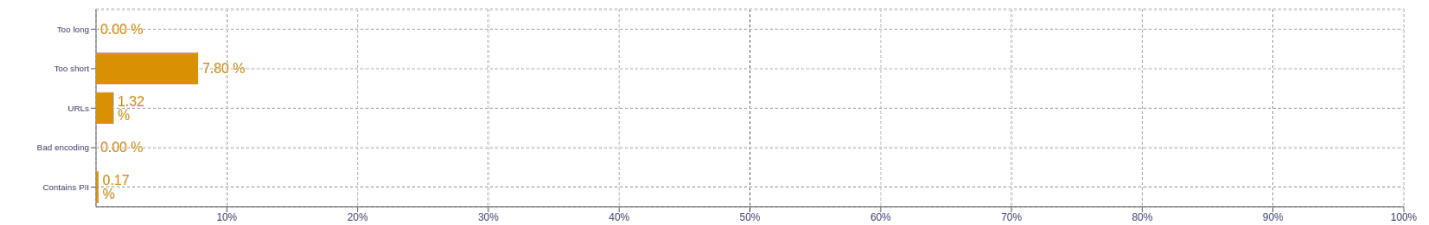
About HPLT Analytics

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