HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-docslite.ru.tsv	7/15/2024	Russian (ru)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
224,196,344	29,428,335,872			3.2 TB	

Top 10 domains

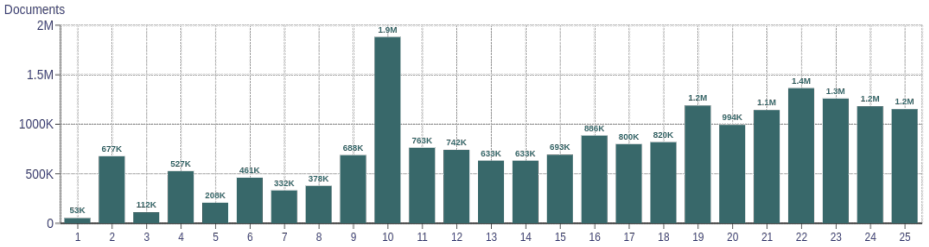
Domain	Docs	% of total
patents.su	1.7M	0.74
blogspot.ru	1.6M	0.72
resheniya-sudov.ru	1.3M	0.56
100book.ru	815K	0.36
resheniya-sudov4.ru	684K	0.30
900-999.ru	681K	0.30
mp3koka.com	572K	0.26
spb.ru	464K	0.21
samlib.ru	456K	0.20
alibaba.com	443K	0.20

Top 10 TLDs

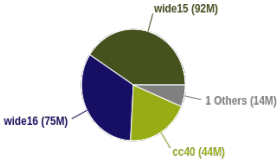
Domain	Docs	% of total
ru	144M	64.08
com	24M	10.83
info	9.3M	4.15
net	7.5M	3.35
org	4.8M	2.15
su	4.5M	2.02
com.ua	4.4M	1.97
by	2.3M	1.04
club	2M	0.87
ua	1.7M	0.77

Documents size (in segments)

<= 25 segments **8.74%** (20M documents)
> 25 segments **91.26%** (204M documents)

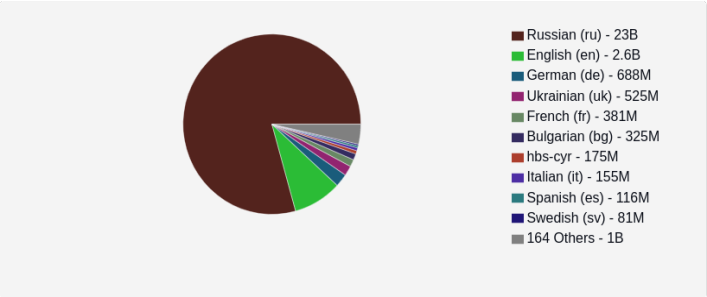


Documents by collection

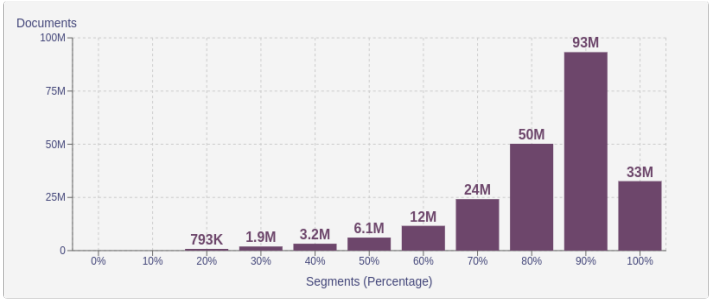


Language Distribution

Number of segments

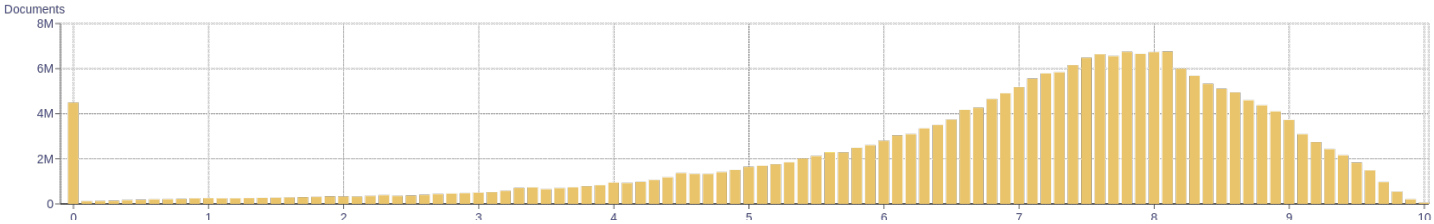


Percentage of segments in Russian (ru) inside documents



Distribution of documents by document score

score <= 5 - **14.04%** (31M documents)
score > 5 - **85.96%** (193M documents)



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