HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-v2-bul_Cyrl.tsv	9/17/2024	Bulgarian (bg)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
28,087,181	681,405,632			159.9 GB	96,280,932,664

Top 10 domains

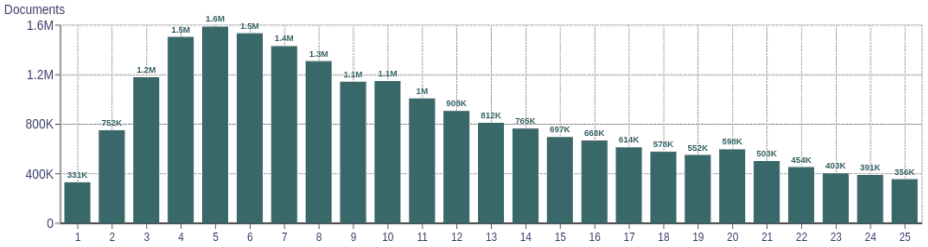
Domain	Docs	% of total
wikipedia.org	553K	1.97
blogspot.com	434K	1.55
grad.bg	347K	1.23
bg-mamma.com	229K	0.82
utre.bg	218K	0.77
gotvach.bg	202K	0.72
wordpress.com	149K	0.53
blog.bg	144K	0.51
dir.bg	139K	0.49
agoda.com	132K	0.47

Top 10 TLDs

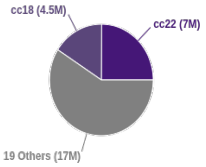
Domain	Docs	% of total
bg	12M	43.25
com	10M	36.05
net	1.5M	5.45
org	1.5M	5.16
eu	949K	3.38
info	774K	2.76
ru	101K	0.36
news	92K	0.33
biz	63K	0.22
de	52K	0.19

Documents size (in segments)

<= 25 segments **75.59%** (21M documents)
> 25 segments **24.41%** (6.9M documents)

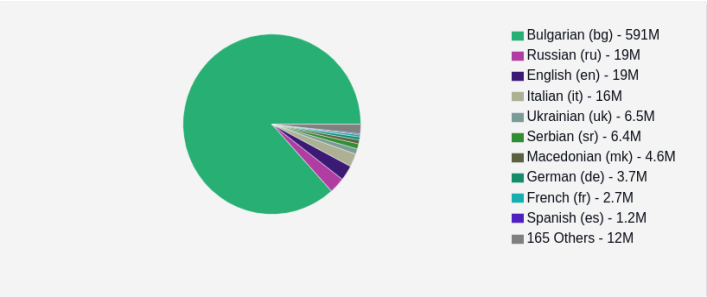


Documents by collection

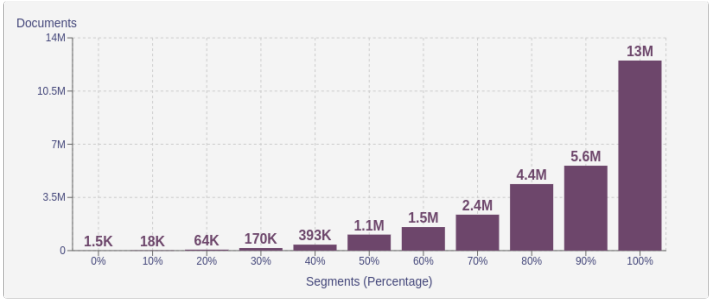


Language Distribution

Number of segments

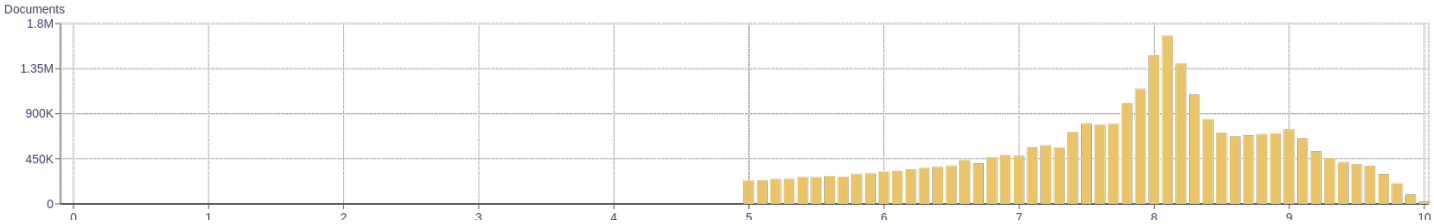


Percentage of segments in Bulgarian (bg) inside documents

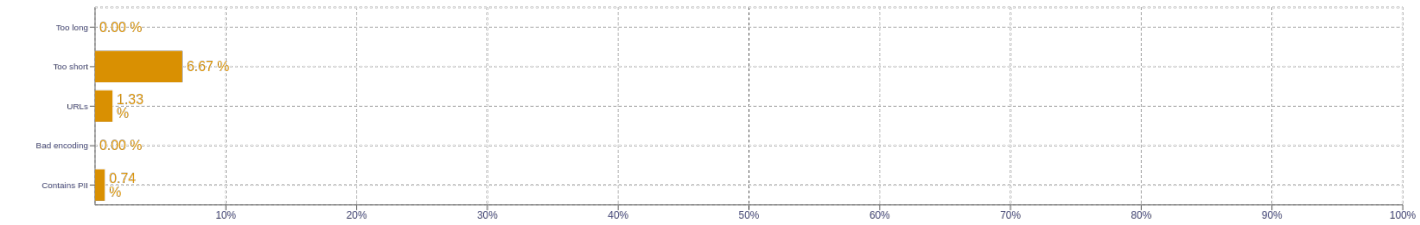


Distribution of documents by document score

score <= 5 - **99.92%** (28M documents)
score > 5 - **0.08%** (23K documents)



Segment noise distribution



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