HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
bul_Cyrl.jsonl.tsv	6/16/2025	Bulgarian (bg)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
28,087,179	681,190,396	275,549,866 (40.45 %)	18B	96,280,932,664	159.9 GB

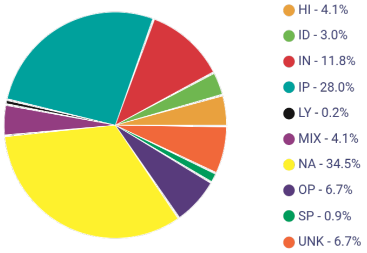
Top 10 domains

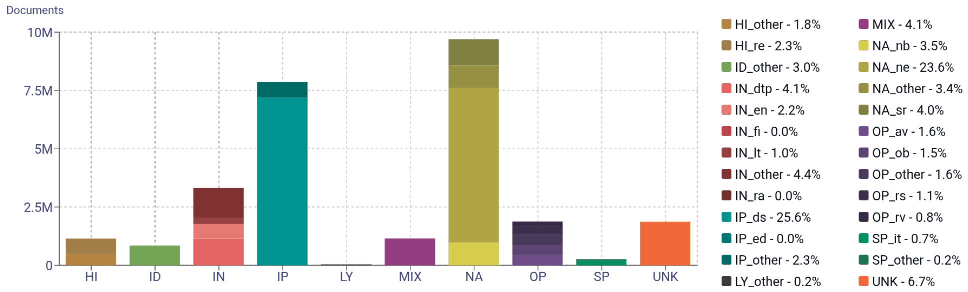
Domain	Docs	% of total
wikipedia.org	553K	1.97%
blogspot.com	434K	1.55%
grad.bg	347K	1.23%
bg-mamma.com	229K	0.82%
utre.bg	218K	0.77%
gotvach.bg	202K	0.72%
wordpress.com	149K	0.53%
blog.bg	144K	0.51%
dir.bg	139K	0.49%
agoda.com	132K	0.47%

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%

Register labels

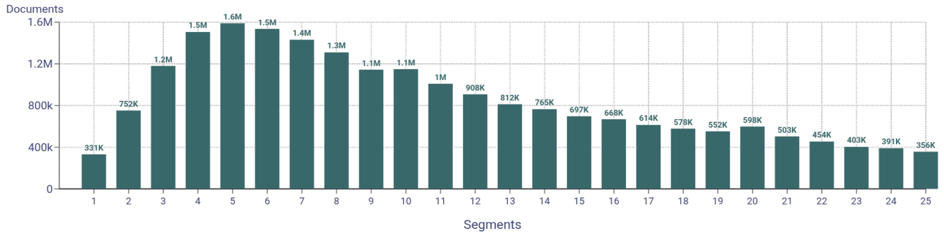




🤖 **MT**:2.9% | 818K Documents

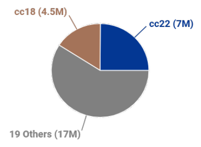
Documents size (in segments)

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



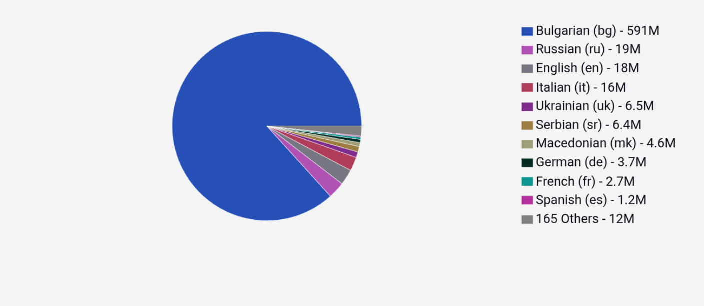
Documents by collection

CC = 59.16%
IA = 40.84%

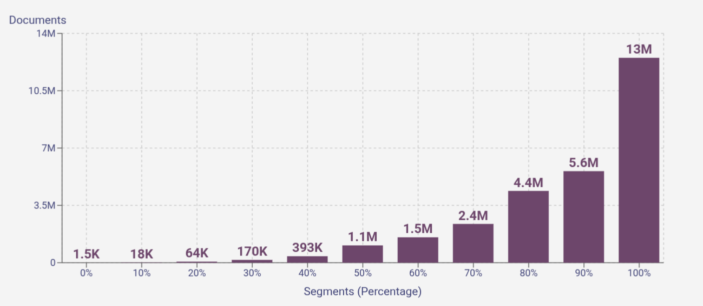


Language Distribution

Number of segments in the Bulgarian (bg) corpus

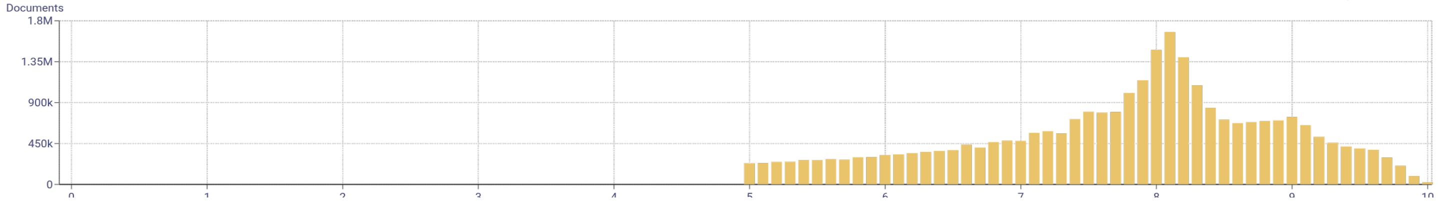


Percentage of segments in Bulgarian (bg) inside documents



Distribution of documents by document score

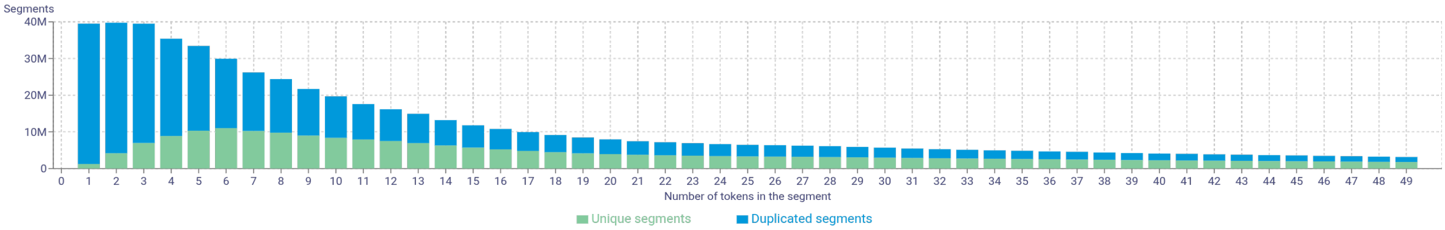
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (28M documents)



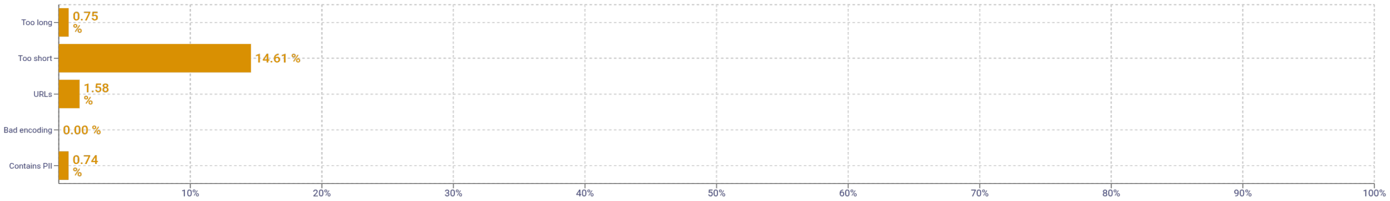
Segment length distribution by token

≤ **49** tokens = **214M** segments | **360M** duplicates

> **50** tokens = **107M** segments | **46M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	г. | 30550952 лв | 16174734 българия | 16116074 част | 13014648 хора | 11539077
2	т. н. | 2019147 т. е. | 1632156 стара загора | 1298746 околната среда | 1022799 крайна сметка | 1006760
3	редактиране на кода | 1280668 начин на приготвяне | 321558 втората световна война | 303775 кирил и методий | 278765 плодове и зеленчуци | 253540
4	защита на личните данни | 237458 опазване на околната среда | 140197 околната среда и водите | 122376 български език и литература | 109992 парламент и на съвета | 87202
5	европейския парламент и на съвета | 85221 оттук гостите имат лесен достъп | 80169 оживен град може да предложи | 79565 wifi достъп във всички стаи | 77256 изменение и допълнение на закона | 74459

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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av