HPLT Analytics report

General overview

Corpus	Date	Language
srp_Cyrl.jsonl.tsv	6/4/2025	Serbian (sr)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
4,123,458	93,792,964	45,007,790 (47.99 %)	2.9B	16,067,193,042	26.72 GB

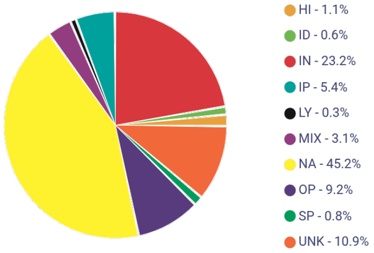
Top 10 domains

Domain	Docs	% of total
wikipedia.org	408K	9.90%
spc.rs	64K	1.56%
iskra.co	62K	1.50%
sputniknews.com	61K	1.47%
vostok.rs	57K	1.39%
srbin.info	56K	1.35%
blogspot.com	53K	1.28%
wordpress.com	50K	1.22%
rbth.com	47K	1.14%
politika.rs	46K	1.13%

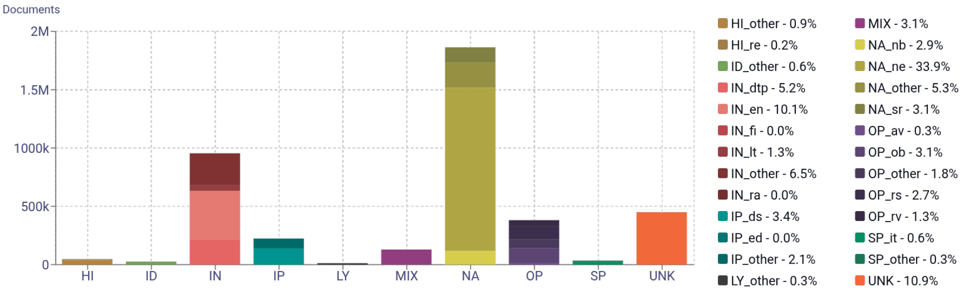
Top 10 TLDs

Domain	Docs	% of total
rs	1.1M	27.81%
com	1M	25.11%
org	707K	17.14%
net	172K	4.18%
org.rs	169K	4.09%
info	157K	3.81%
gov.rs	120K	2.91%
edu.rs	117K	2.85%
co.rs	68K	1.65%
co	64K	1.54%

Register labels

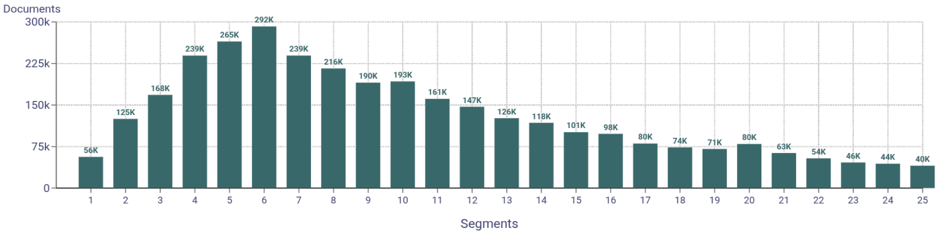


🤖 **MT**:7.8% | 321K Documents



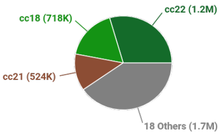
Documents size (in segments)

<= 25 segments **79.67%** (3.3M documents)
> 25 segments **20.33%** (838K documents)



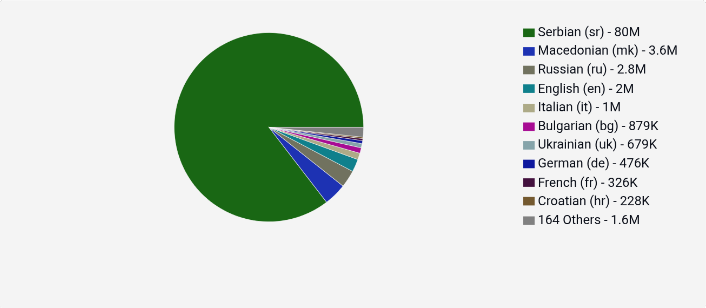
Documents by collection

CC = 69.74%
IA = 30.26%

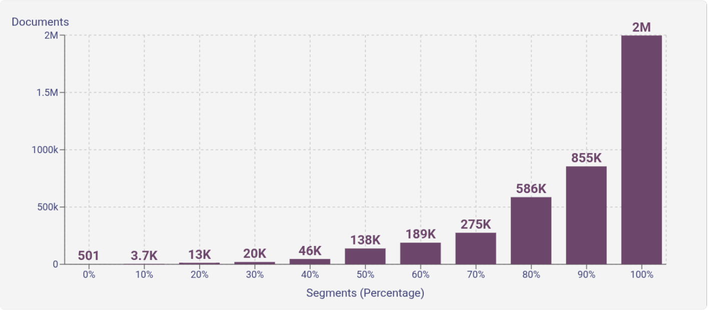


Language Distribution

Number of segments in the Serbian (sr) corpus

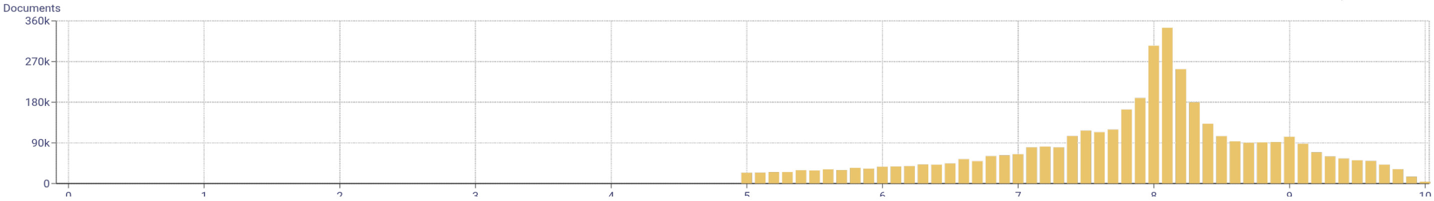


Percentage of segments in Serbian (sr) inside documents



Distribution of documents by document score

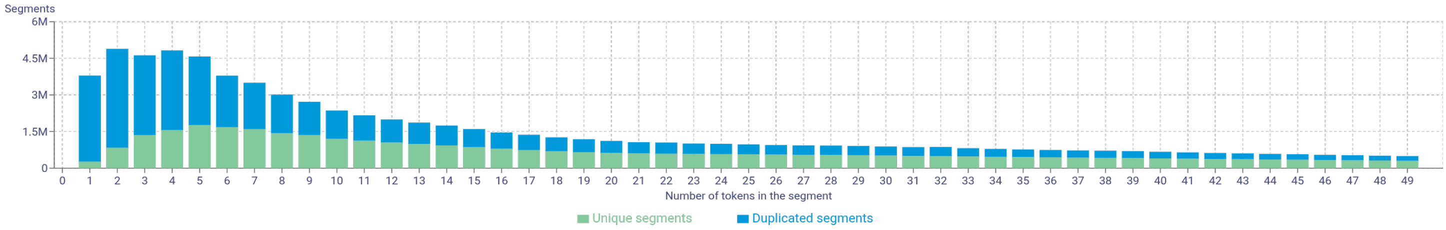
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (4.1M documents)



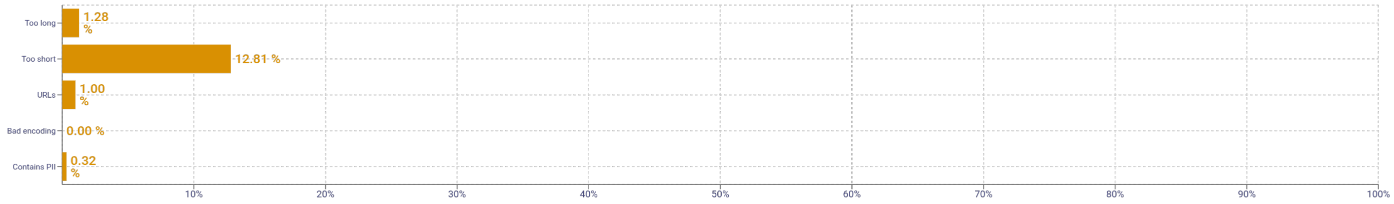
Segment length distribution by token

≤ **49** tokens = **35M** segments | **41M** duplicates

> **50** tokens = **18M** segments | **7.3M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	године | 7426709 није | 5737621 све | 5175484 како | 4907490 која | 4810313
2	републике српске | 522810 републике србије | 464157 због тога | 431624 пре свега | 328231 без обзира | 273966
3	другог светског рата | 149172 косову и метохији | 142677 косова и метохије | 127448 српске православне цркве | 117824 босне и херцеговине | 110143
4	osigurajte pravo na profil | 57229 науке и технолошког развоја | 50077 бнио бнио бнио бнио | 40688 стижу директно на вашу | 35289 директно на вашу e-mail | 35289
5	архивирано из оригинала на датум | 49229 бнио бнио бнио бнио бнио | 40635 стижу директно на вашу e-mail | 35289 директно на вашу e-mail адресу | 35289 vlasnik ovog objekta ili upravljate | 28615

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