HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
bos_Latn.jsonl.tsv	6/9/2025	Bosnian (bs)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
14,613,084	268,122,004	121,540,640 (45.33 %)	8.4B	45,817,404,394	44.01 GB

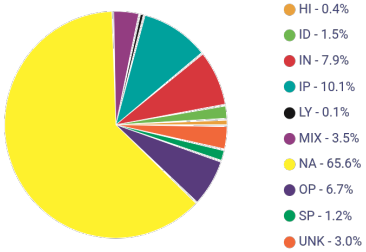
Top 10 domains

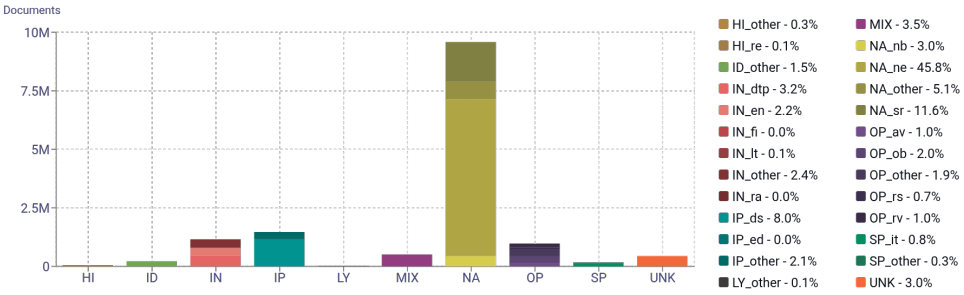
Domain	Docs	% of total
klix.ba	372K	2.55%
wikipedia.org	273K	1.87%
sportske.net	184K	1.26%
vesti.rs	181K	1.24%
slobodnaevropa.org	174K	1.19%
blogspot.com	155K	1.06%
blic.rs	140K	0.96%
krstarica.com	113K	0.78%
b92.net	108K	0.74%
mondo.rs	105K	0.72%

Top 10 TLDs

Domain	Docs	% of total
com	4.4M	29.93%
rs	3.9M	26.89%
ba	2M	13.45%
net	1.3M	8.78%
org	891K	6.10%
info	501K	3.43%
me	461K	3.16%
hr	195K	1.33%
org.rs	164K	1.12%
co.rs	141K	0.96%

Register labels

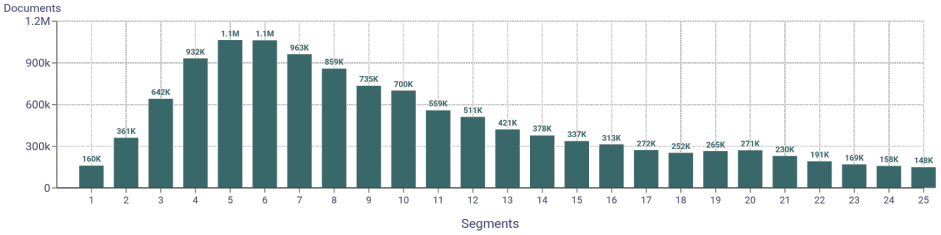




MT:0.4% | 64K Documents

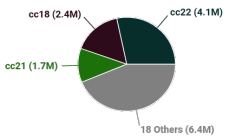
Documents size (in segments)

<= 25 segments **81.79%** (12M documents)
> 25 segments **18.21%** (2.7M documents)



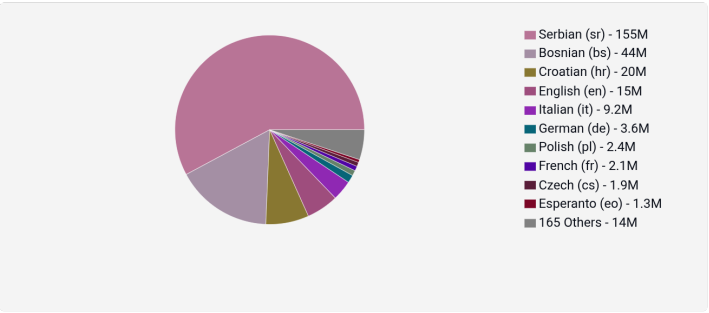
Documents by collection

CC = 67.63%
IA = 32.37%

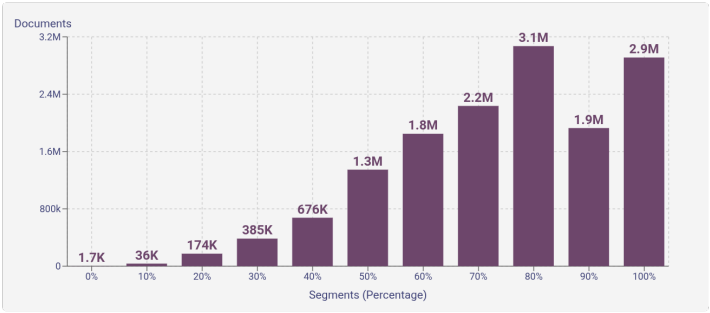


Language Distribution

Number of segments in the Bosnian (bs) corpus

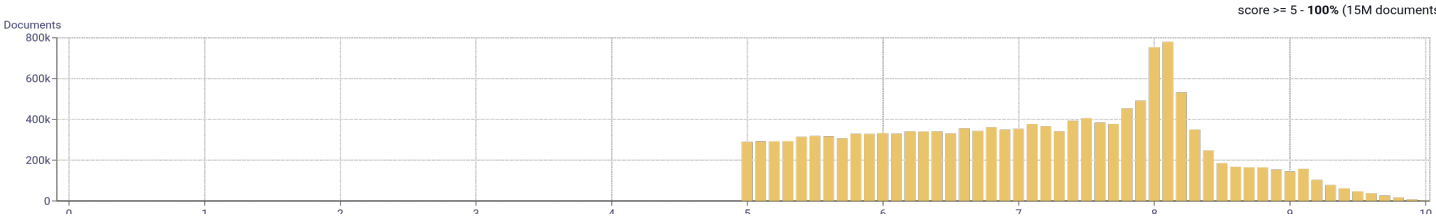


Percentage of segments in Bosnian (bs) inside documents



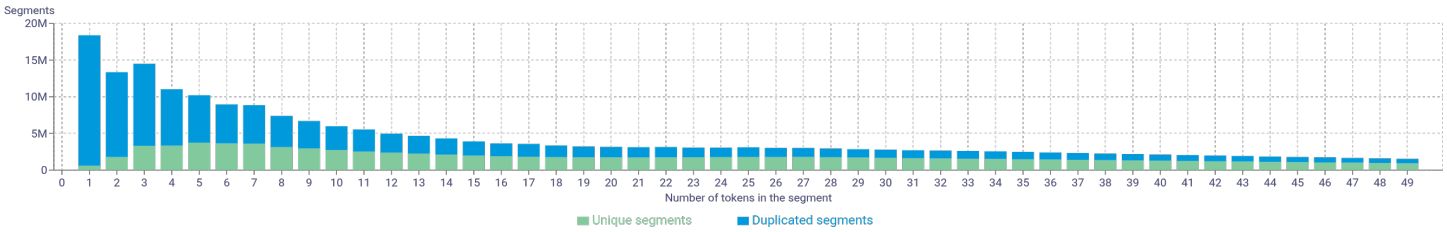
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (15M documents)

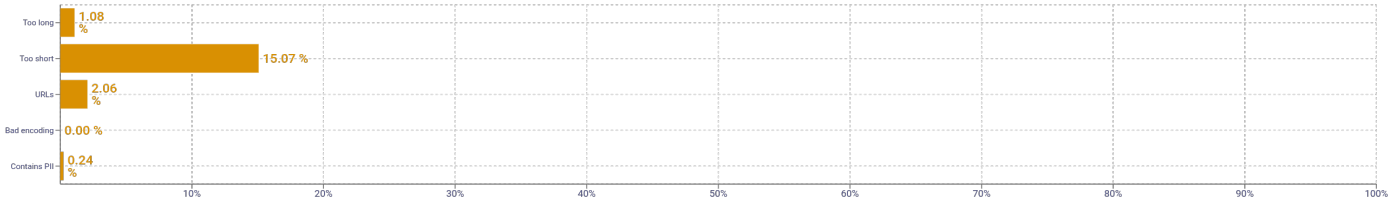


Segment length distribution by token

≤ 49 tokens = **91M** segments | **125M** duplicates
> 50 tokens = **52M** segments | **21M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	godine | 19141983 · godina | 8266428 · dana | 8245316 · srbije | 7407131 · dok | 7207789
2	prvi put | 1305580 · prošle godine | 991093 · republike srpske | 931610 · druge strane | 890202 · crnoj gori | 782843
3	bosne i hercegovine | 1287837 · bosni i hercegovini | 707641 · bosna i hercegovina | 305285 · alejhi ve sellem | 209065 · drugog svetskog rata | 191481
4	sallallahu alejhi ve sellem | 178379 · navodi se u saopštenju | 143766 · s. a. v. s. | 123035 · celu vest na sajtu | 117694 · nalazi se u ulici | 101171
5	pročitaj celu vest na sajtu | 117456 · komentarima su privatno mišljenje autora | 84493 · iznešena u komentarima su privatno | 78717 · komentara i ne odražavaju stavove | 76694 · autora komentara i ne odražavaju | 76694

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