HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
HPLT-v2-bos_Latn.tsv	9/22/2024	Bosnian (bs)

Volumes

Docs	Segments	Characters	Size
14,613,088	268,156,601	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

- HI - 0.4%
- ID - 1.5%
- IN - 7.9%
- IP - 10.1%
- LY - 0.1%
- MIX - 3.5%
- NA - 65.6%
- OP - 6.7%
- SP - 1.2%
- UNK - 3.0%

MT:0.4% | 64K Documents

- HI_other - 0.3%
- HI_re - 0.1%
- ID_other - 1.5%
- IN_dtp - 3.2%
- IN_en - 2.2%
- IN_fi - 0.0%
- IN_lt - 0.1%
- IN_other - 2.4%
- IN_ra - 0.0%
- IP_ds - 8.0%
- IP_ed - 0.0%
- IP_other - 2.1%
- LY_other - 0.1%
- MIX - 3.5%
- NA_nb - 3.0%
- NA_ne - 45.8%
- NA_other - 5.1%
- NA_sr - 11.6%
- OP_av - 1.0%
- OP_ob - 2.0%
- OP_other - 1.9%
- OP_rs - 0.7%
- OP_rv - 1.0%
- SP_it - 0.8%
- SP_other - 0.3%
- UNK - 3.0%
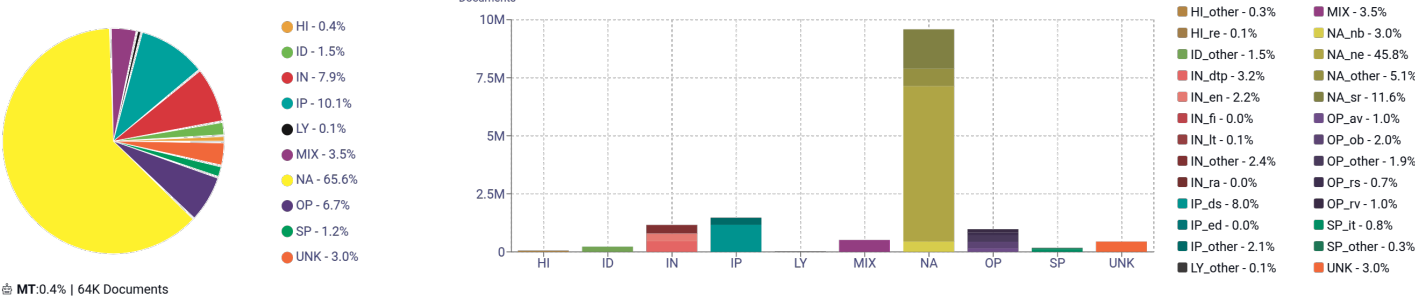
Documents size (in segments)

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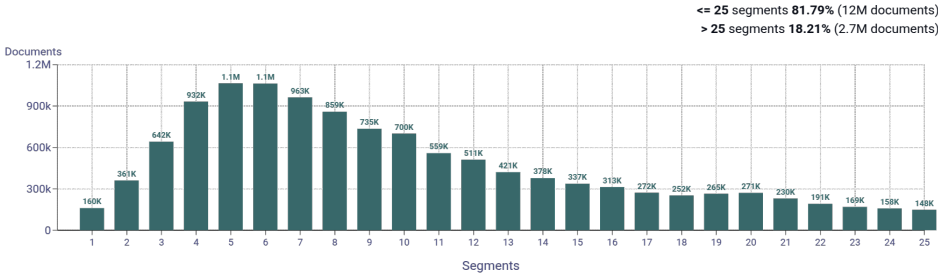
Documents by collection

CC = 67.63%
IA = 32.37%

- cc18 (2.4M)
- cc21 (1.7M)
- cc22 (4.1M)
- 18 Others (6.4M)
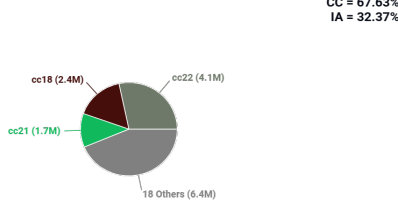
Language Distribution

Number of segments in the Bosnian (bs) corpus

- Bosnian (bs) - 100M
- Serbian (sr) - 88M
- Croatian (hr) - 31M
- English (en) - 15M
- Italian (it) - 9.2M
- German (de) - 3.6M
- Polish (pl) - 2.4M
- French (fr) - 2.1M
- Czech (cs) - 1.9M
- Slovenian (sl) - 1.4M
- 165 Others - 14M
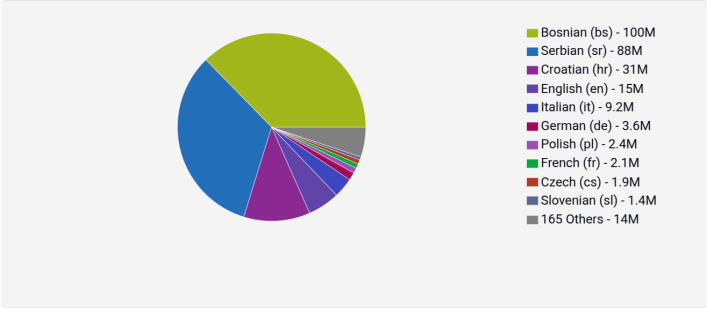
### Percentage of segments in Bosnian (bs) inside documents
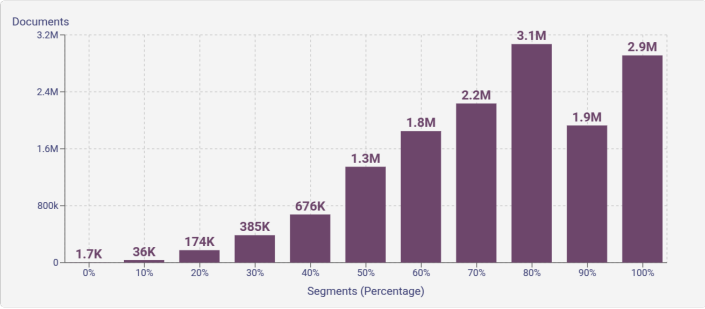
Distribution of documents by document score

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

Segment noise distribution

- Too long: 1.08 %
- Too short: 15.08 %
- URLs: 2.06 %
- Bad encoding: 0.00 %
- Contains PII: 0.24 %
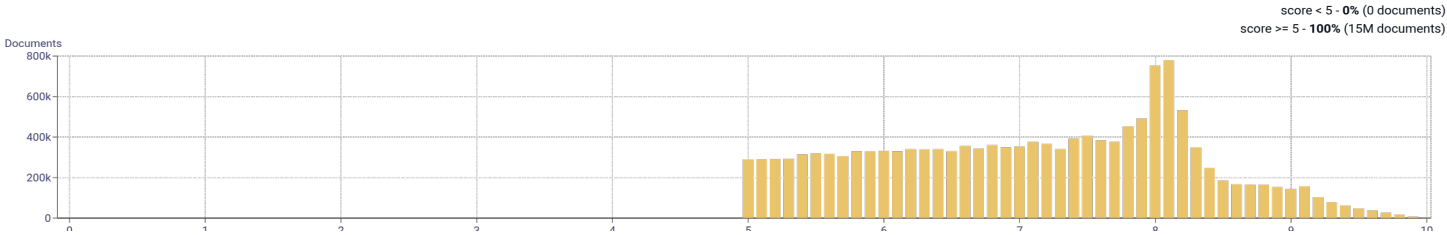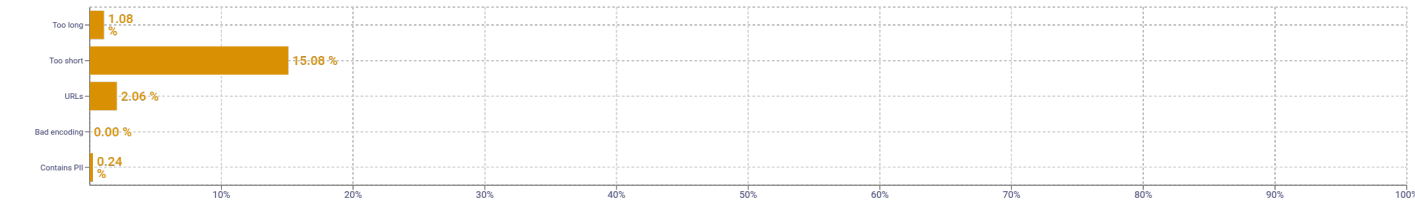
About HPLT Analytics

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.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	How-to or instructions	HI	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	Informational persuasion	IP	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	Opinion	OP
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	Informational description	IN	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				