HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
HPLT-v2-pol_Latn.tsv	9/7/2024	Polish (pl)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
175,410,669	4,460,824,697			615.5 GB	627,308,403,809

Top 10 domains

Domain	Docs	% of total
blogspot.com	3.7M	2.12
onet.pl	1.8M	1.05
wikipedia.org	1.5M	0.84
wp.pl	1M	0.59
gazeta.pl	992K	0.57
interia.pl	923K	0.53
sfd.pl	761K	0.43
wordpress.com	722K	0.41
docplayer.pl	620K	0.35
frazeo.pl	440K	0.25

Top 10 TLDs

Domain	Docs	% of total
pl	124M	70.75
com	17M	9.54
com.pl	6.2M	3.53
org	4.2M	2.37
eu	4M	2.26
net	2.7M	1.56
info	2.2M	1.27
org.pl	1.6M	0.91
edu.pl	1.5M	0.85
net.pl	1.2M	0.68

Documents size (in segments)

<= 25 segments **73.01%** (128M documents)
> 25 segments **26.99%** (47M documents)
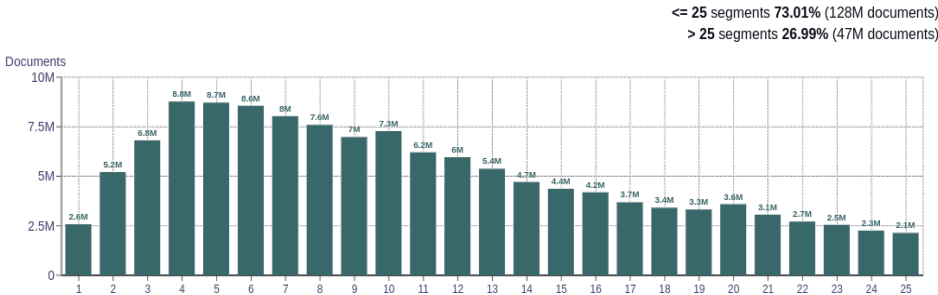
Documents by collection

cc18 (29M)
cc22 (38M)
19 Others (109M)
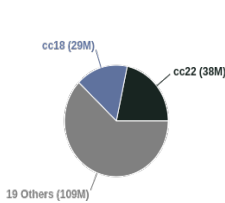
Language Distribution

Number of segments

- Polish (Latin) (pol-Latn) - 3.4B
- Enawené-Nawé (unk) - 256M
- szl (Latin) (szl-Latn) - 222M
- English (Latin) (eng-Latn) - 80M
- Slovenian (Latin) (slv-Latn) - 21M
- Bosnian (Latin) (bos-Latn) - 18M
- Norwegian Nynorsk (Latin) (nno-Latn) - 18M
- vec (Latin) (vec-Latn) - 14M
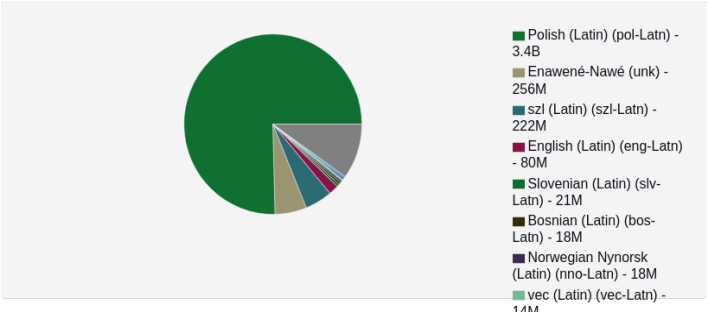
Percentage of segments in Polish (pl) inside documents
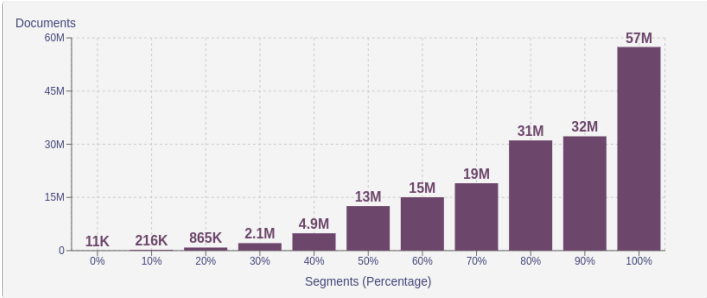
Segments (Percentage)	Documents
0%	11K
10%	216K
20%	865K
30%	2.1M
40%	4.9M
50%	13M
60%	15M
70%	19M
80%	31M
90%	32M
100%	57M

Distribution of documents by document score

score <= 5 - **99.94%** (175M documents)
score > 5 - **0.06%** (107K documents)

Segment noise distribution

Contains PII — 0.56%
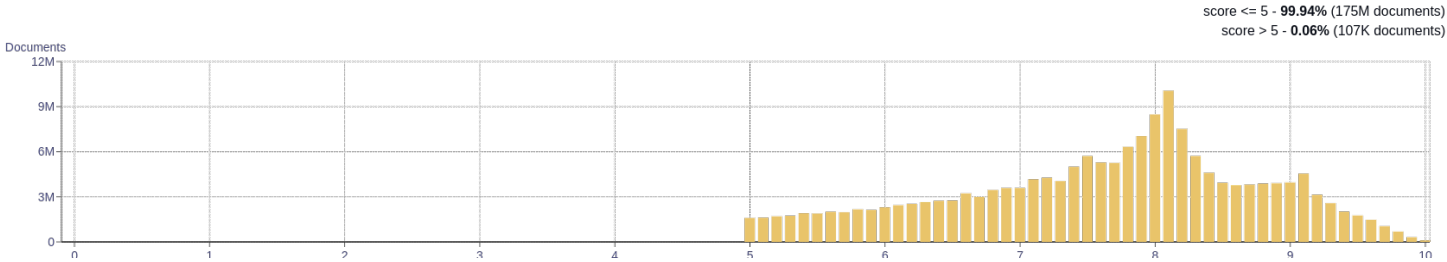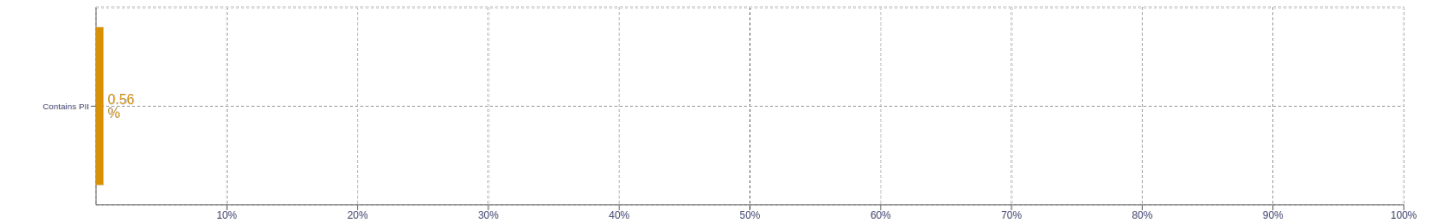
About HPLT Analytics

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