HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
epo_Latn.jsonl.tsv	9/16/2024	Esperanto (eo)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
818,878	20,353,314	7,149,533 (35.13 %)	571M	2.81 GB	2,956,534,128

Top 10 domains

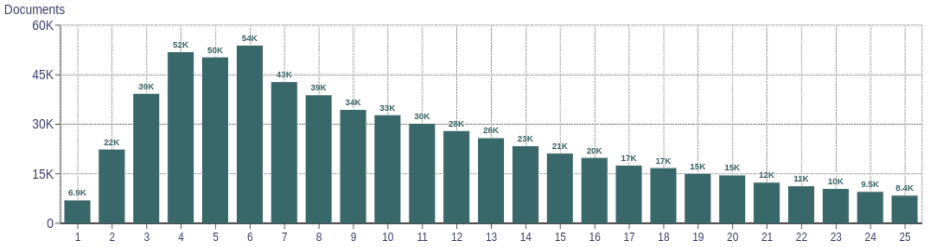
Domain	Docs	% of total
wikipedia.org	538K	65.69
blogspot.com	9.9K	1.21
wikitrans.net	8.3K	1.02
esperantio.net	7.2K	0.88
pola-retradio.org	5.7K	0.70
wordpress.com	5.7K	0.70
espero.com.cn	4.6K	0.57
over-blog.com	3.8K	0.46
ikso.net	3.6K	0.44
uea.org	3.6K	0.44

Top 10 TLDs

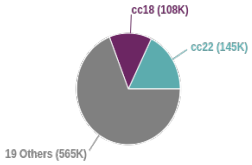
Domain	Docs	% of total
org	595K	72.69
com	99K	12.15
net	39K	4.74
ru	10K	1.22
cn	6.5K	0.80
info	6.5K	0.79
be	5.2K	0.63
com.cn	4.7K	0.57
de	4.6K	0.56
eu	4.5K	0.55

Documents size (in segments)

<= 25 segments **77.76%** (637K documents)
> 25 segments **22.24%** (182K documents)

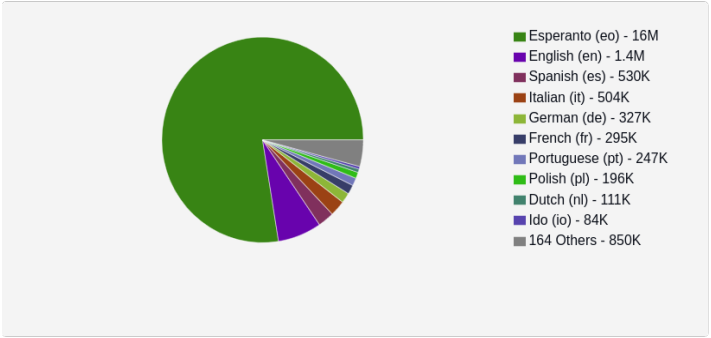


Documents by collection

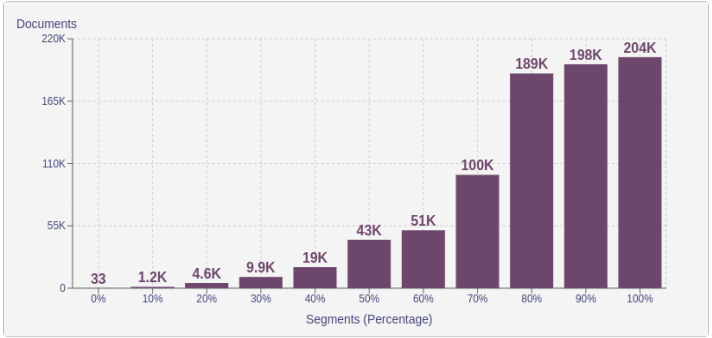


Language Distribution

Number of segments

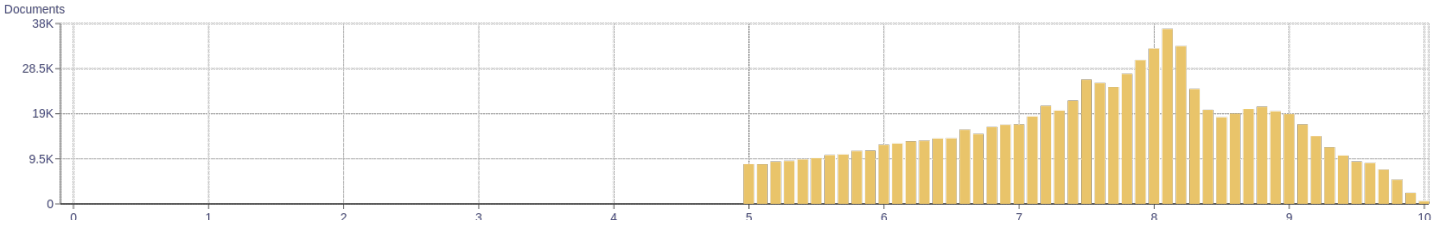


Percentage of segments in Esperanto (eo) inside documents



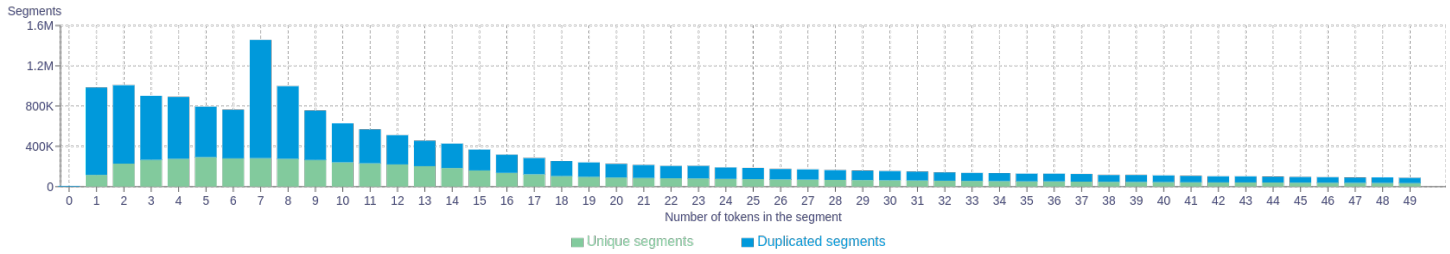
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (819K documents)

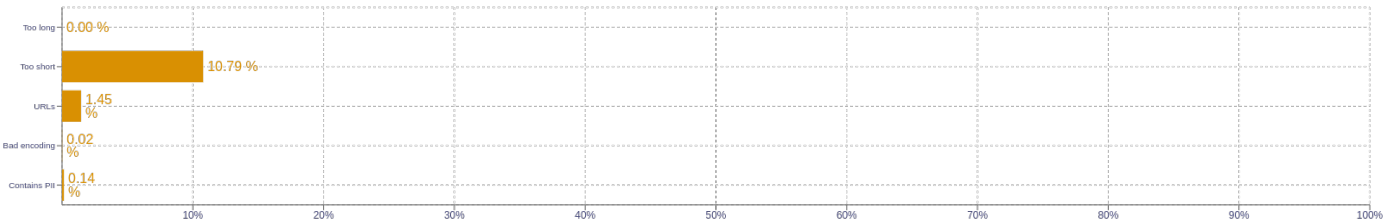


Segment length distribution by token

<= 49 tokens = **5.7M** segments | **11M** duplicates
> 50 tokens = **3.5M** segments | **2.1M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	redakti | 3013130 · kun | 2377814 · pri | 2275296 · kiel | 2189232 · el | 2157729
2	redakti fonton | 1405770 · povas esti | 202460 · of the | 118446 · eksteraj ligiloj | 113827 · temas pri | 108717
3	iom post iom | 30695 · per la retarkivo | 26652 · retarkivo wayback machine | 26472 · el la jaro | 25792 · ekde la jaro | 17807
4	per la retarkivo wayback | 26569 · arkivita el la originalo | 14587 · archived from the original | 11946 · from the original on | 11183 · el la plej gravaj | 10845
5	per la retarkivo wayback machine | 26472 · archived from the original on | 11114 · ankaŭ en la vikimedia komunejo | 8898 · vidu ankaŭ en la vikimedia | 8880 · kolekto de bildoj kaj plurmediaj | 8852

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