HPLT Analytics report

General overview

Corpus	Analytics date	Language
HPLT-v2-lvs_Latn.tsv	9/29/2024	Latvian (lvs)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
6,771,634	173,807,290			25.39 GB	25,015,494,704

Top 10 domains

Domain	Docs	% of total
wikipedia.org	253K	3.74
atlants.lv	199K	2.93
tvnet.lv	122K	1.80
skaties.lv	93K	1.37
hotels.com	89K	1.32
viss.lv	85K	1.26
delfi.lv	74K	1.09
lsm.lv	72K	1.06
agoda.com	61K	0.90
blogspot.com	54K	0.79

Top 10 TLDs

Domain	Docs	% of total
lv	4.8M	70.33
com	1.1M	16.32
org	327K	4.82
gov.lv	126K	1.86
eu	122K	1.80
info	57K	0.84
net	45K	0.66
ie	19K	0.27
edu.lv	17K	0.24
lt	16K	0.23

Documents size (in segments)

<= 25 segments **77.02%** (5.2M documents)
> 25 segments **22.98%** (1.6M documents)
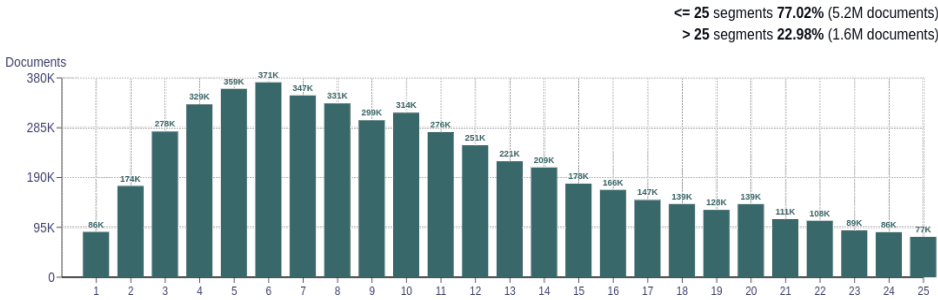
Documents by collection

cc18 (1.2M) cc22 (1.7M) cc21 (888K) 18 Others (3M)
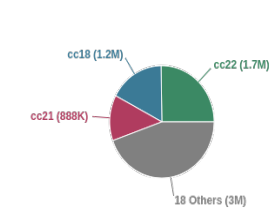
Language Distribution

Number of segments

- Latvian (lv) - 145M
- English (en) - 7.2M
- Italian (it) - 4.4M
- French (fr) - 2M
- German (de) - 1.6M
- Lithuanian (lt) - 1.5M
- Dutch (nl) - 1.5M
- Spanish (es) - 1M
- Russian (ru) - 883K
- Finnish (fi) - 835K
- 165 Others - 8M

*Latvian (lvs) identification might be inaccurate because language is not supported by Fasttext
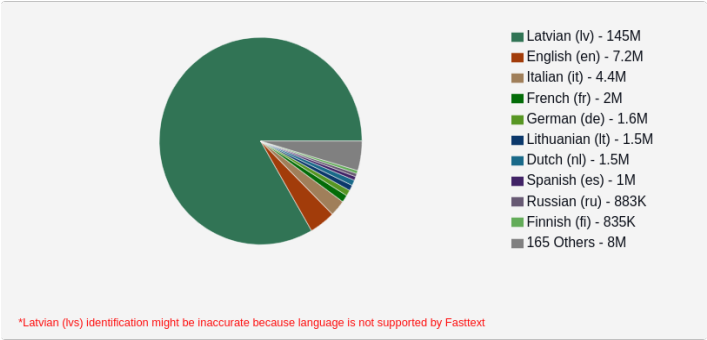
Percentage of segments in Latvian (lvs) inside documents
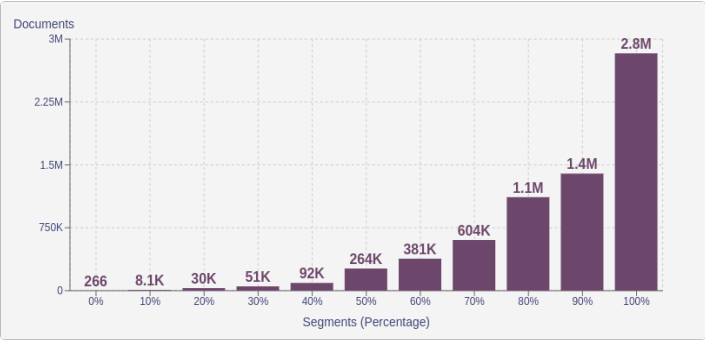
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (6.8M documents)

Segment noise distribution

Too long	0.65 %
Too short	15.44 %
URLs	3.82 %
Bad encoding	0.00 %
Contains PII	0.37 %
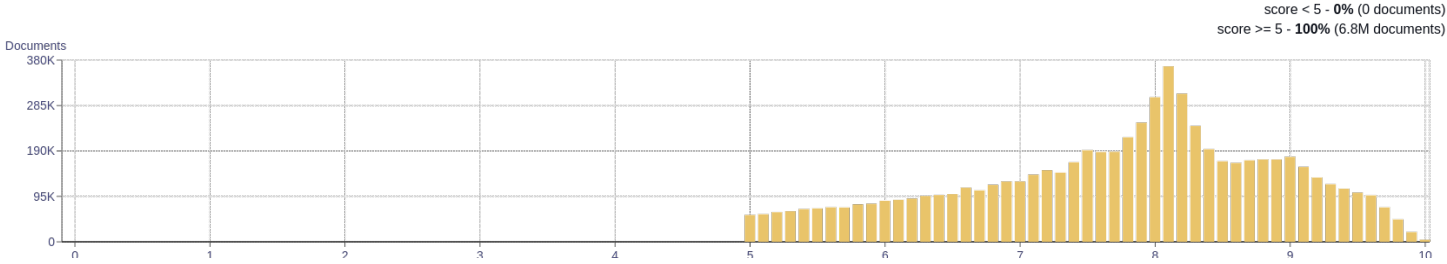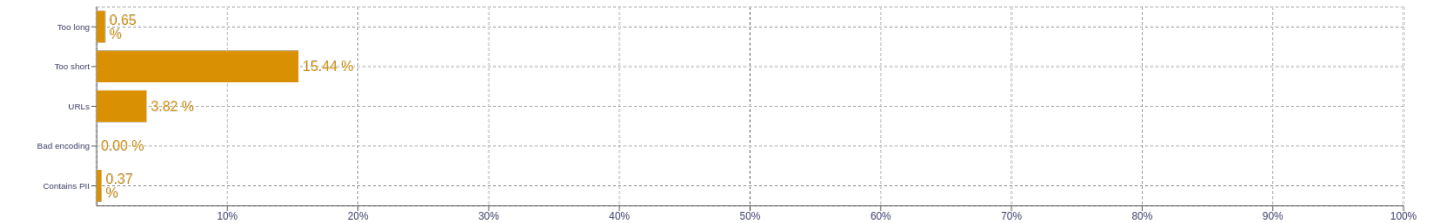
About HPLT Analytics

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