HPLT Analytics report

General overview

Corpus	Analytics date	Language
HPLT-v2-vie_Latn.tsv	9/25/2024	Vietnamese (vi)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
100,746,136	3,019,551,654			457.87 GB	376,569,396,268

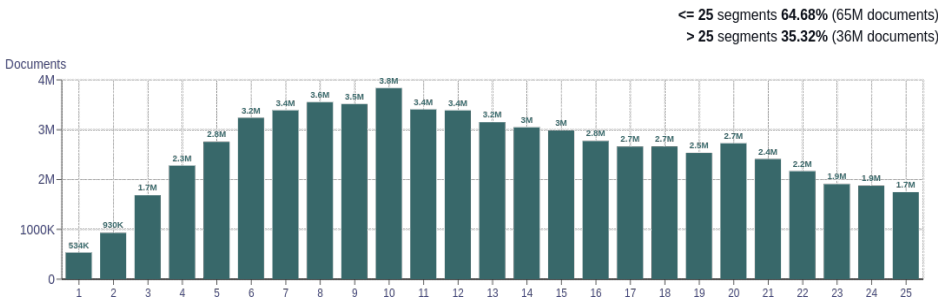### Top 10 domains

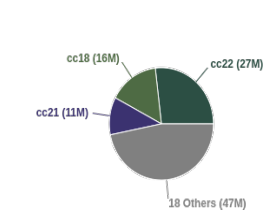| Domain	Docs	% of total
blogspot.com	921K	0.91
vietbao.vn	762K	0.76
wordpress.com	748K	0.74
wikipedia.org	559K	0.56
tin247.com	493K	0.49
vnexpress.net	452K	0.45
tailieu.vn	424K	0.42
baomoi.com	365K	0.36
vtv.vn	359K	0.36
nld.com.vn	355K	0.35

Top 10 TLDs

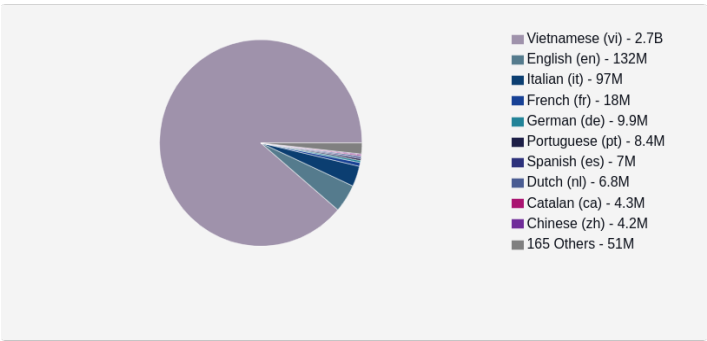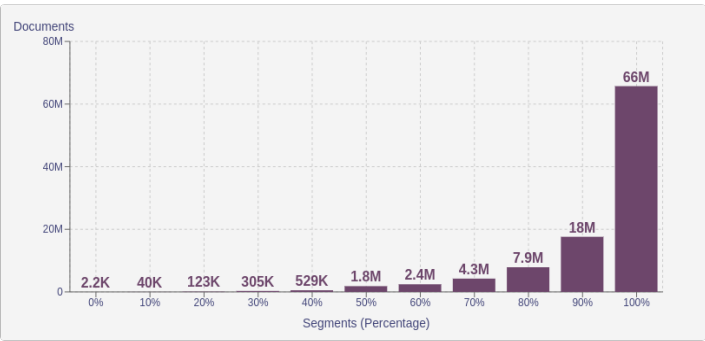| Domain	Docs	% of total
com	38M	37.72
vn	30M	30.14
com.vn	8.6M	8.57
net	8.3M	8.28
org	3.6M	3.59
edu.vn	2.6M	2.60
info	1.4M	1.43
gov.vn	1.1M	1.14
org.vn	878K	0.87
net.vn	709K	0.70

Documents size (in segments)

<= 25 segments **64.68%** (65M documents)
> 25 segments **35.32%** (36M documents)

Documents by collection

Language Distribution

Number of segments

- Vietnamese (vi) - 2.7B
- English (en) - 132M
- Italian (it) - 97M
- French (fr) - 18M
- German (de) - 9.9M
- Portuguese (pt) - 8.4M
- Spanish (es) - 7M
- Dutch (nl) - 6.8M
- Catalan (ca) - 4.3M
- Chinese (zh) - 4.2M
- 165 Others - 51M

Percentage of segments in Vietnamese (vi) inside documents

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (101M documents)

Segment noise distribution

- Too long: 0.00 %
- Too short: 8.19 %
- URLs: 1.64 %
- Bad encoding: 0.11 %
- Contains PII: 0.26 %
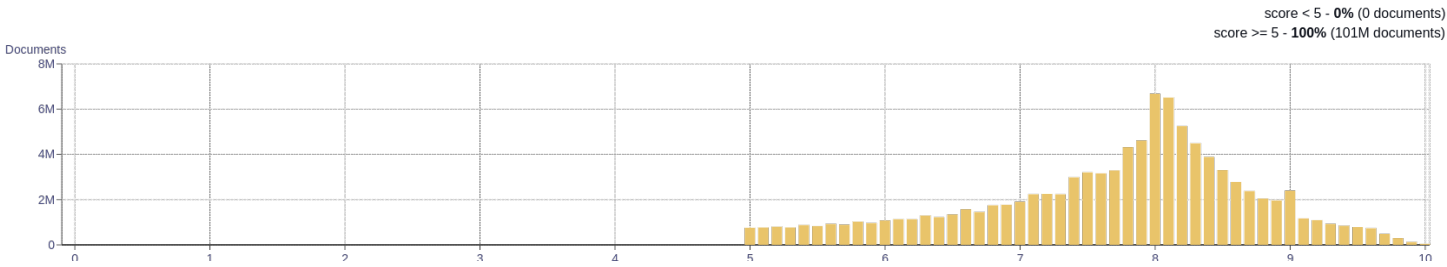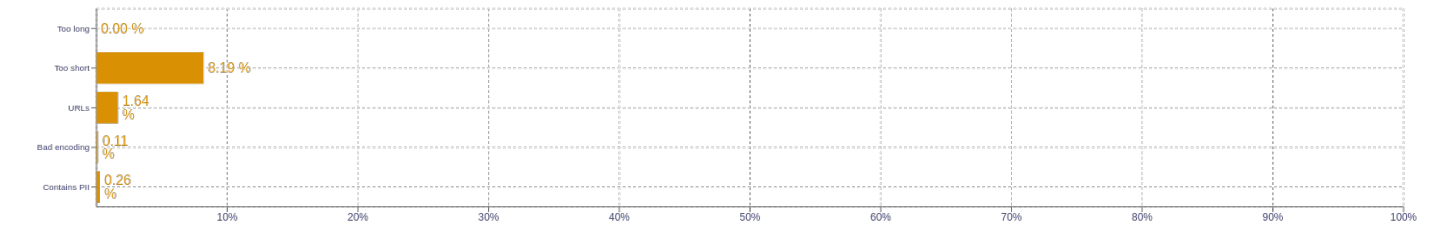
About HPLT Analytics

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