HPLT Analytics report

General overview

Corpus	Analytics date	Language
HPLT-docslite.vi.tsv	6/28/2024	Vietnamese (vi)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
31,504,473	4,650,885,367			274.62 GB	

Top 10 domains

Domain	Docs	% of total
alibaba.com	1.1M	3.37
aliexpress.com	253K	0.80
blogspot.com	201K	0.64
vatgia.com	185K	0.59
wikidich.com	176K	0.56
dailyvina.com	145K	0.46
muasimviettel.com	123K	0.39
diebuchsuche.com	111K	0.35
vnexpress.net	100K	0.32
simsodep.pro.vn	86K	0.27

Top 10 TLDs

Domain	Docs	% of total
com	15M	48.60
vn	6.2M	19.73
net	2.4M	7.57
com.vn	2.2M	6.93
org	1.1M	3.53
edu.vn	915K	2.90
info	511K	1.62
gov.vn	273K	0.87
org.vn	204K	0.65
net.vn	187K	0.59

Documents size (in segments)

<= 25 segments **6.94%** (2.2M documents)
> 25 segments **93.06%** (29M documents)
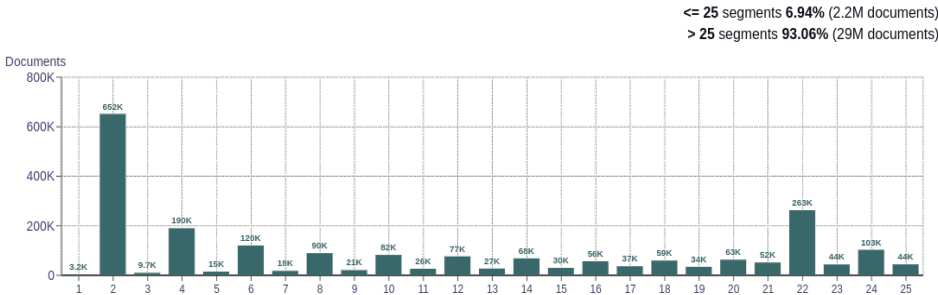
Documents by collection

cc40 (11M), wide16 (7.6M), wide17 (6.1M), wide15 (6.8M)
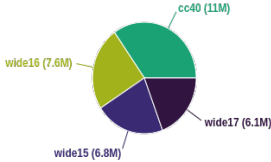
Language Distribution

Number of segments

- Vietnamese (vi) - 3.4B
- English (en) - 738M
- French (fr) - 71M
- Italian (it) - 55M
- German (de) - 47M
- Spanish (es) - 35M
- Russian (ru) - 26M
- Portuguese (pt) - 22M
- Catalan (ca) - 16M
- Swedish (sv) - 15M
- 164 Others - 182M
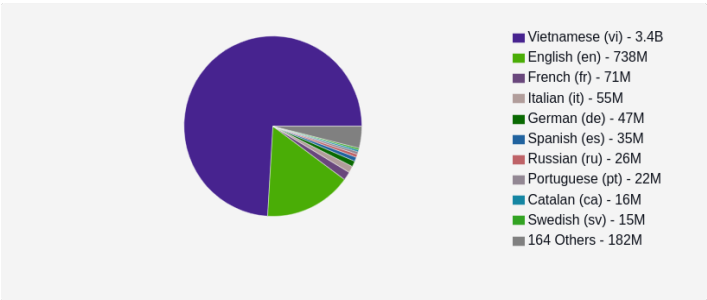
### Percentage of segments in Vietnamese (vi) inside documents

Distribution of documents by document score

score <= 5 - **24.63%** (7.8M documents)
score > 5 - **75.37%** (24M documents)
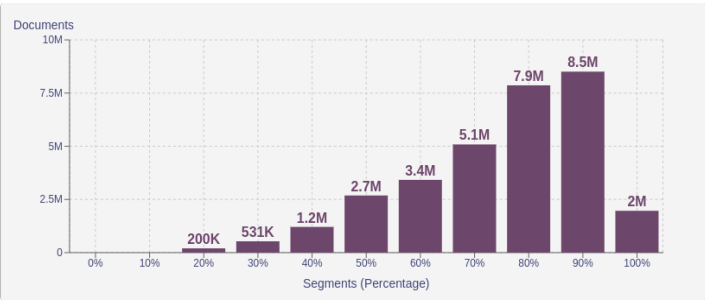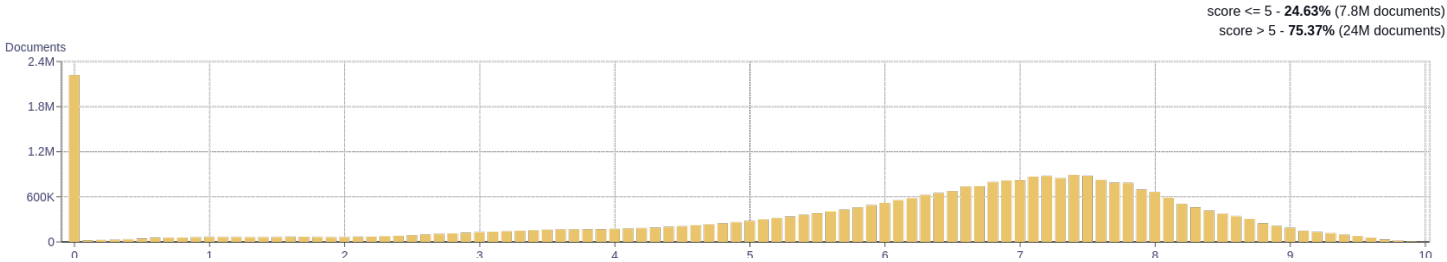
About HPLT Analytics

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