HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
vie_Latn.jsonl.tsv	8/10/2025	Vietnamese (vi)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
100,746,132	3,017,881,429	1,298,427,668 (43.02 %)	94B	376,569,396,268	457.87 GB

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%

Register labels

- HI - 3.8%
- ID - 2.3%
- IN - 13.8%
- IP - 32.1%
- LY - 0.8%
- MIX - 4.0%
- NA - 28.8%
- OP - 6.0%
- SP - 0.4%
- UNK - 8.0%

- HI_other - 3.0%
- HI_re - 0.8%
- ID_other - 2.3%
- IN_dtp - 3.7%
- IN_en - 0.7%
- IN_fi - 0.0%
- IN_lt - 0.8%
- IN_other - 8.6%
- IN_ra - 0.0%
- IP_ds - 29.5%
- IP_ed - 0.0%
- IP_other - 2.6%
- LY_other - 0.8%
- MIX - 4.0%
- NA_nb - 1.4%
- NA_ne - 19.2%
- NA_other - 5.4%
- NA_sr - 2.7%
- OP_av - 2.0%
- OP_ob - 0.7%
- OP_other - 1.3%
- OP_rs - 1.1%
- OP_rv - 0.8%
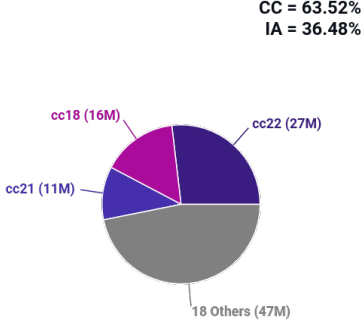- SP_it - 0.2%
- SP_other - 0.2%
- UNK - 8.0%
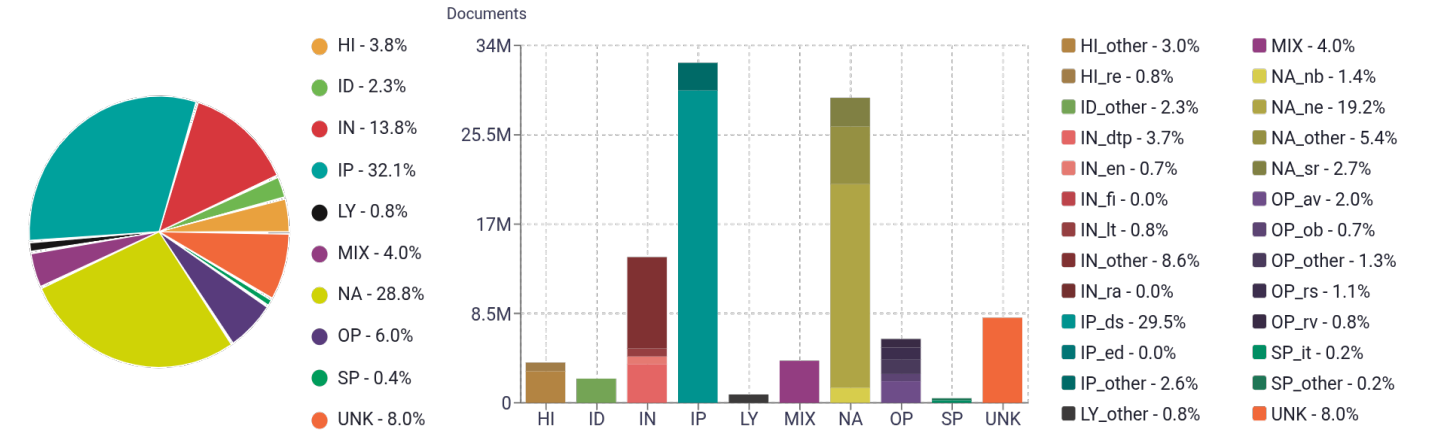
🤖 **MT**:3.1% | 3.1M Documents

Documents size (in segments) ⓘ
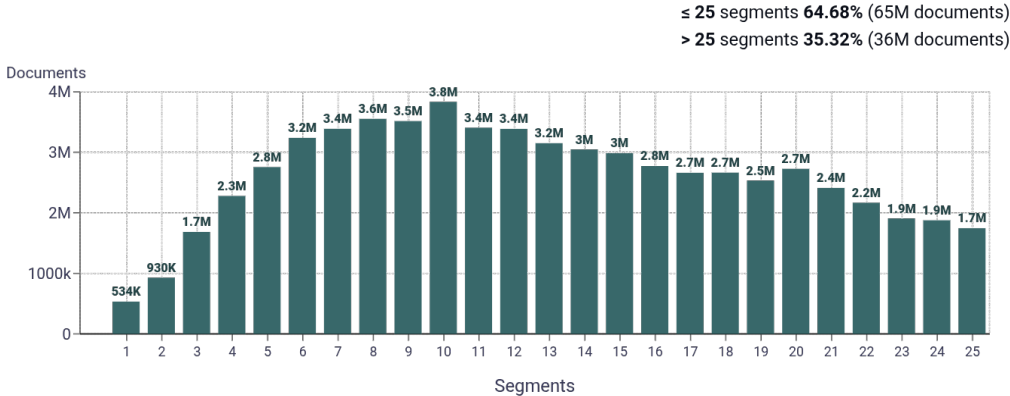
≤ 25 segments **64.68%** (65M documents)
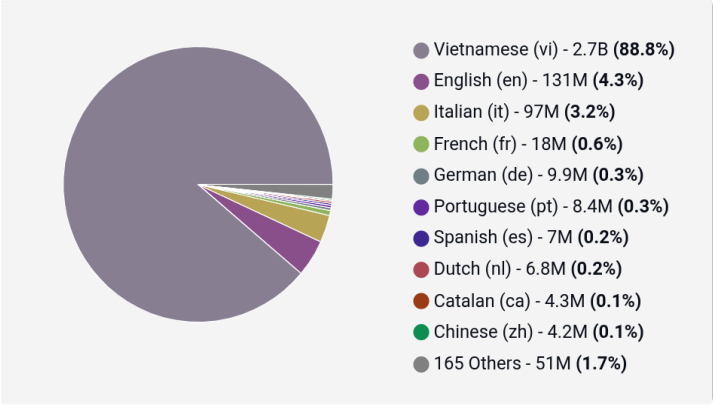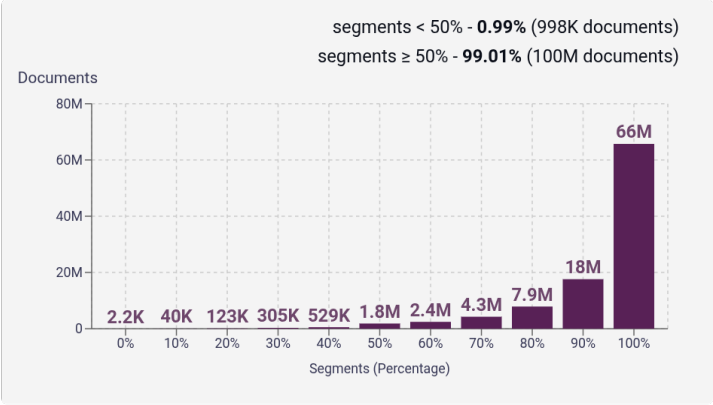> 25 segments **35.32%** (36M documents)
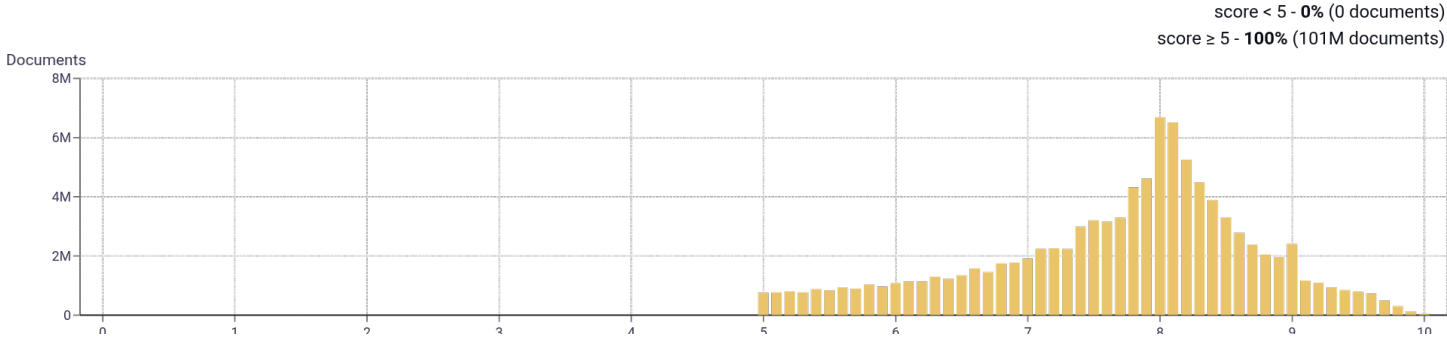

Document collections

CC = 63.52%
IA = 36.48%

cc18 (16M)
cc21 (11M)
cc22 (27M)
18 Others (47M)

Language Distribution

Number of segments in the Vietnamese (vi) corpus

- Vietnamese (vi) - 2.7B **(88.8%)**
- English (en) - 131M **(4.3%)**
- Italian (it) - 97M **(3.2%)**
- French (fr) - 18M **(0.6%)**
- German (de) - 9.9M **(0.3%)**
- Portuguese (pt) - 8.4M **(0.3%)**
- Spanish (es) - 7M **(0.2%)**
- Dutch (nl) - 6.8M **(0.2%)**
- Catalan (ca) - 4.3M **(0.1%)**
- Chinese (zh) - 4.2M **(0.1%)**
- 165 Others - 51M **(1.7%)**

Percentage of segments in Vietnamese (vi) inside documents

segments < 50% - **0.99%** (998K documents)
segments ≥ 50% - **99.01%** (100M documents)

Distribution of documents by document score

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

Segment length distribution by token

≤ **49** tokens = **2.4B** segments | **1.5B** duplicates
> **50** tokens = **614M** segments | **260M** duplicates

Segment noise distribution

- Too long — 0.37%
- Too short — 9.51%
- URLs — 1.82%
- Bad encoding — 0.11%
- Contains PII — 0.26%

■ % of corpus
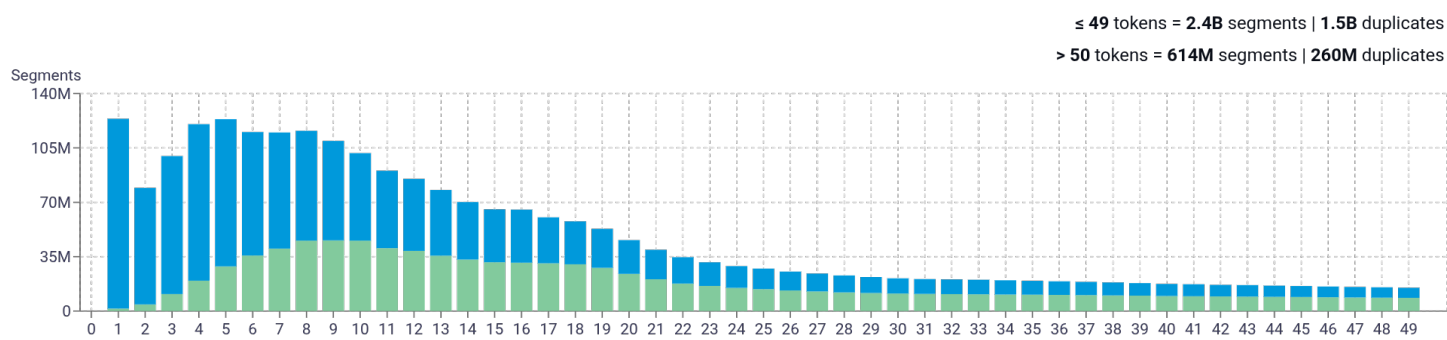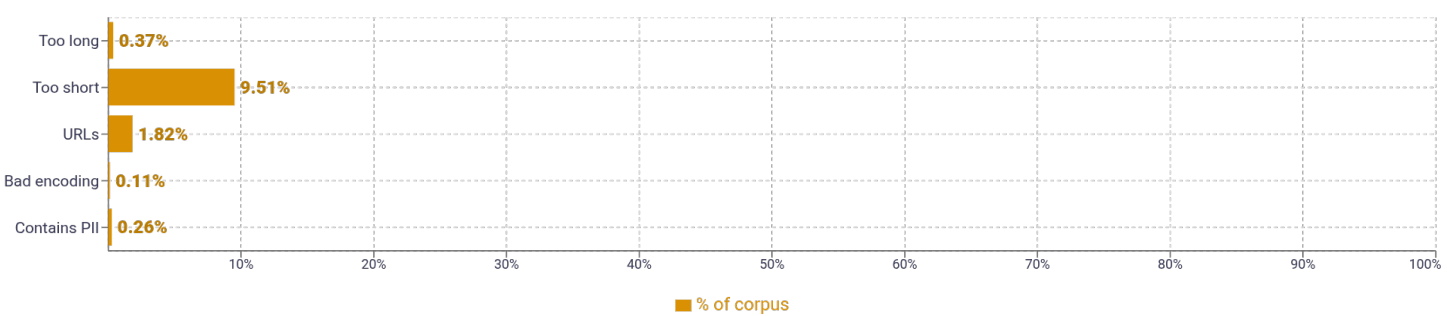
Frequent n-grams

SIZE	N-GRAMS	
1	công | 370,935,826 thể | 370,152,961 hàng | 241,681,382 dụng | 228,454,409 thành | 226,850,121	⧉
2	sử dụng | 125,727,789 việt nam | 95,931,943 sản phẩm | 93,579,558 công ty | 73,377,078 thời gian | 70,901,128	⧉
3	hồ chí minh | 14,821,213 bất động sản | 8,231,186 công ty tnhh | 8,123,791 trường đại học | 6,938,498 quý khách hàng | 6,775,489	⧉
4	công ty cổ phần | 6,241,512 cung cấp dịch vụ | 3,213,554 chúng ta có thể | 2,904,628 công nghệ thông tin | 2,011,189 bảo vệ môi trường | 1,858,694	⧉
5	thành phố hồ chí minh | 3,407,881 quy định của pháp luật | 2,306,601 giáo dục và đào tạo | 1,459,712 thị trường bất động sản | 1,192,507 xin vui lòng liên hệ | 950,771	⧉

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

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	HI	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	**Informational persuasion**	IP	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	**Opinion**	OP
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	**Informational description**	IN	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				