

General overview

Corpus	Date	Language
hplt-v3-vie_Latn	9/19/2025	Vietnamese (vi)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
145,403,073	3,590,720,809	2,068,629,153 (57.61 %)	117B	472,178,502,228	573.71 GB

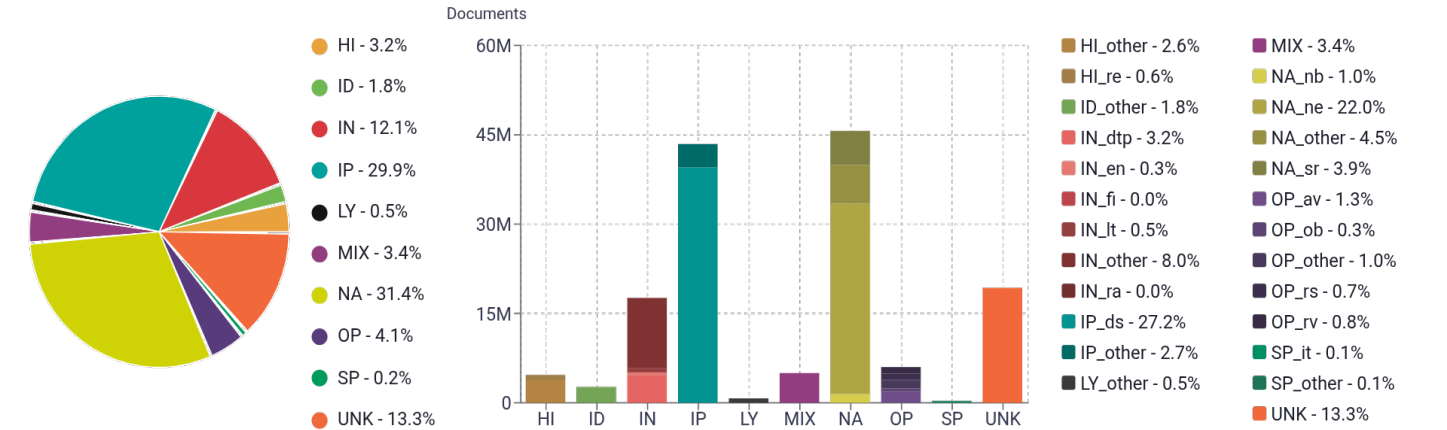
Top 10 domains

Domain	Docs	% of total
tuoitre.vn	846K	0.58%
vietbao.vn	775K	0.53%
wordpress.com	726K	0.50%
blogspot.com	654K	0.45%
tienphong.vn	597K	0.41%
thanhnien.vn	582K	0.40%
vietnamnet.vn	579K	0.40%
vnexpress.net	557K	0.38%
baomoi.com	531K	0.37%
tin247.com	450K	0.31%

Top 10 TLDs

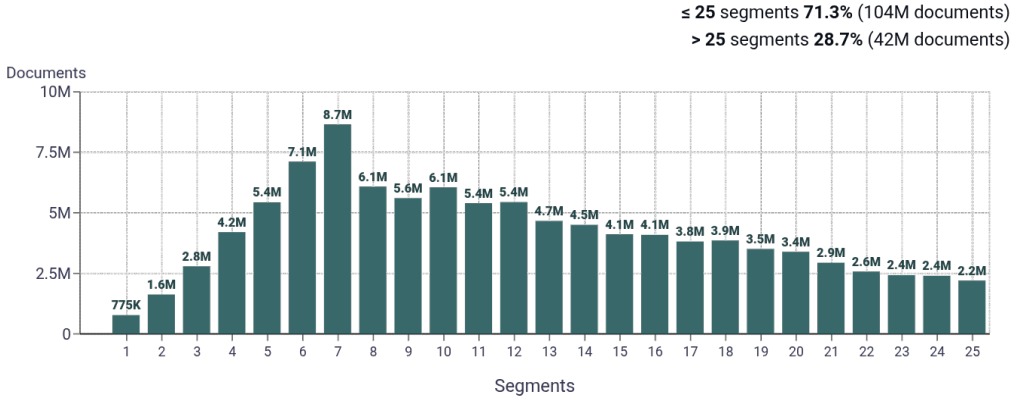
Domain	Docs	% of total
com	54M	37.05%
vn	46M	31.32%
com.vn	12M	7.91%
net	11M	7.30%
edu.vn	4.2M	2.89%
org	3.8M	2.62%
gov.vn	2.4M	1.68%
tw	1.6M	1.13%
info	1.5M	1.02%
org.vn	1.3M	0.87%

Register labels

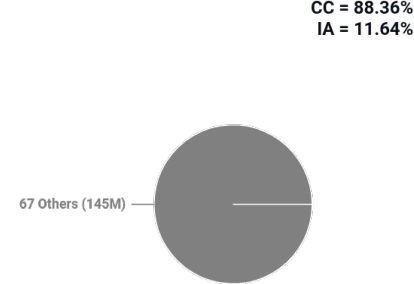


MT:7.9% | 11M Documents

Documents size (in segments) ⓘ

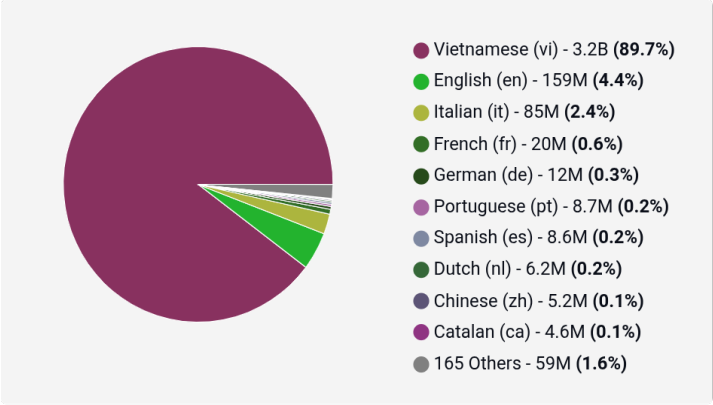


Document collections

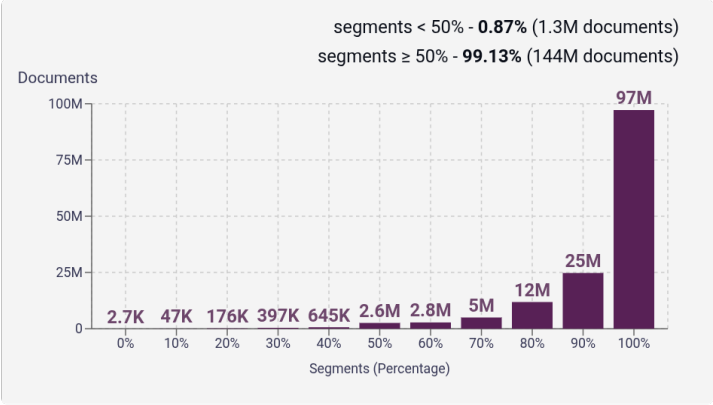


Language Distribution

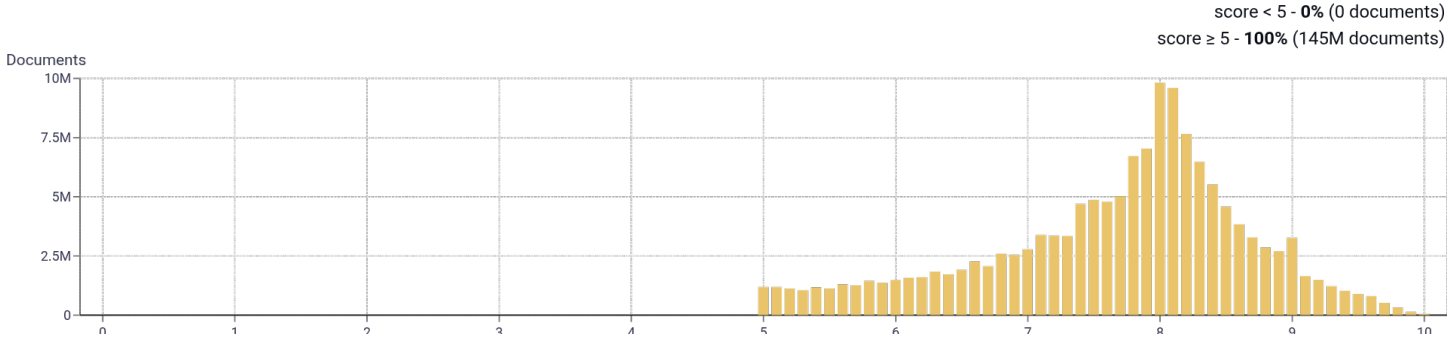
Number of segments in the Vietnamese (vi) corpus



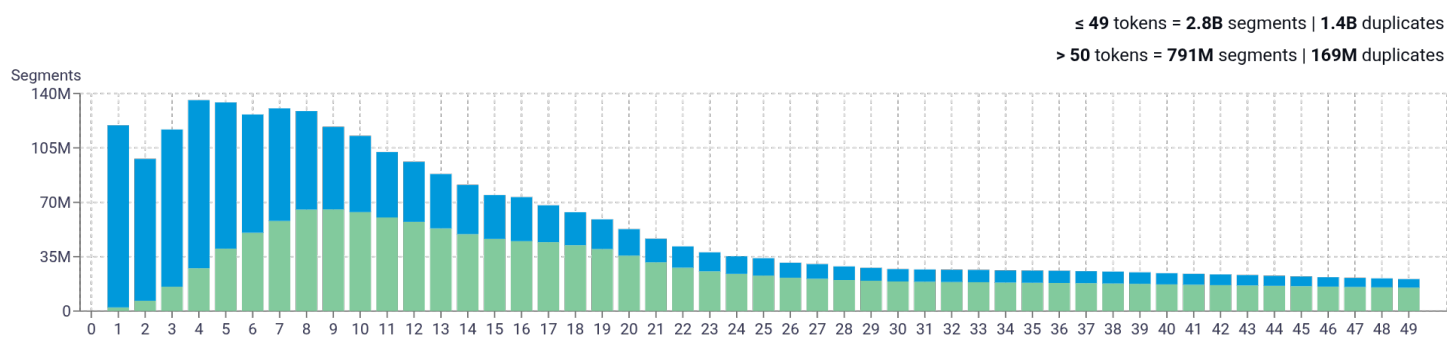
Percentage of segments in Vietnamese (vi) inside documents



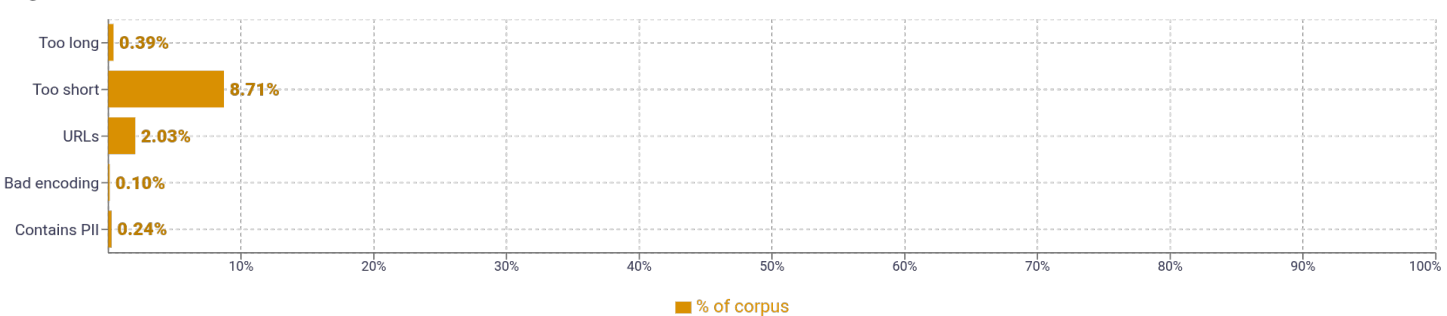
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>công 470,286,446</div> <div>thể 463,445,102</div> <div>hàng 305,925,138</div> <div>dụng 304,529,567</div> <div>thành 286,206,917</div>	
2	<div>sử dụng 168,439,062</div> <div>sản phẩm 126,290,063</div> <div>việt nam 121,748,226</div> <div>thời gian 89,008,541</div> <div>công ty 88,061,506</div>	
3	<div>hồ chí minh 17,783,073</div> <div>công ty TNHH 9,836,215</div> <div>bất động sản 9,407,070</div> <div>trường đại học 8,512,438</div> <div>toàn có thể 8,093,670</div>	
4	<div>hoàn toàn có thể 8,034,279</div> <div>công ty cổ phần 7,454,012</div> <div>cảm ơn tình yêu 5,464,427</div> <div>cá cược bóng đá 5,300,867</div> <div>đảm bảo an toàn 4,931,168</div>	
5	<div>thành phố hồ chí minh 4,455,276</div> <div>quy định của pháp luật 2,663,848</div> <div>giáo dục và đào tạo 1,937,778</div> <div>lệ cá cược bóng đá 1,569,666</div> <div>phù hợp với nhu cầu 1,534,267</div>	

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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>, , , 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 stopwords. 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				