

General overview

Corpus	Date	Language
hplt-v3-eng_Latn	9/24/2025	English (en)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
3,216,000	78,157,600	65,761,988 (84.14 %)	2.5B	12,879,848,191	12.08 GB

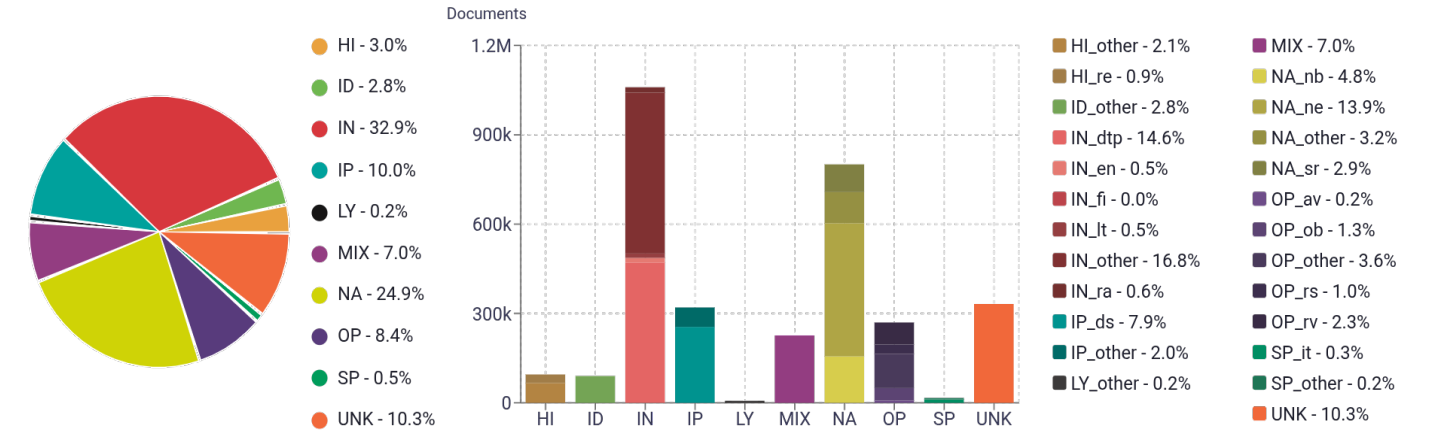
Top 10 domains

Domain	Docs	% of total
blogspot.com	92K	2.86%
wordpress.com	41K	1.26%
typepad.com	4.4K	0.14%
tumblr.com	3.4K	0.11%
stackexchange.com	3.4K	0.10%
wikipedia.org	3K	0.09%
yahoo.com	2.8K	0.09%
cbslocal.com	2.6K	0.08%
google.com	2.5K	0.08%
blogspot.co.uk	2.5K	0.08%

Top 10 TLDs

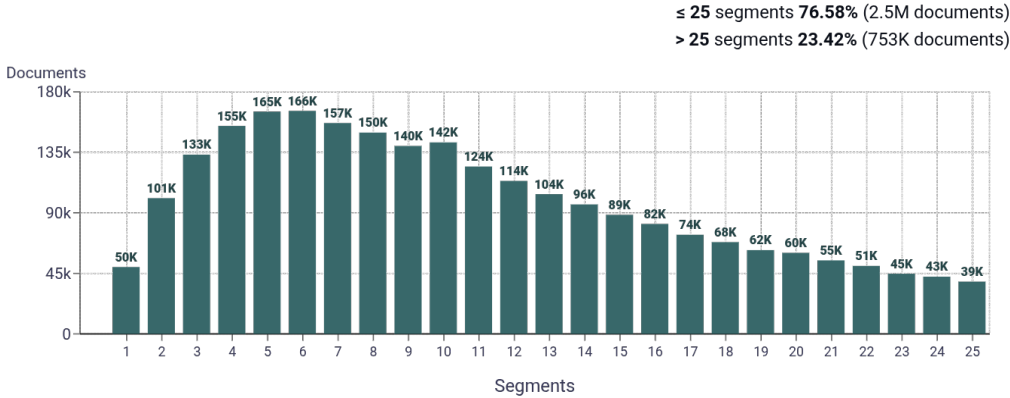
Domain	Docs	% of total
com	2.1M	66.85%
org	269K	8.38%
co.uk	118K	3.66%
net	104K	3.23%
edu	54K	1.69%
com.au	49K	1.52%
ca	41K	1.29%
info	21K	0.66%
in	21K	0.64%
gov	18K	0.57%

Register labels



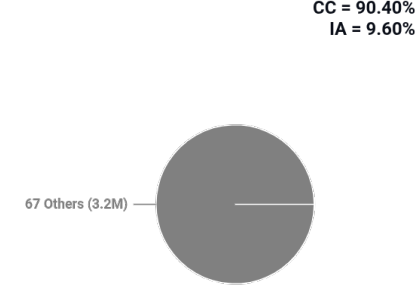
MT:1.9% | 61K Documents

Documents size (in segments) ⓘ



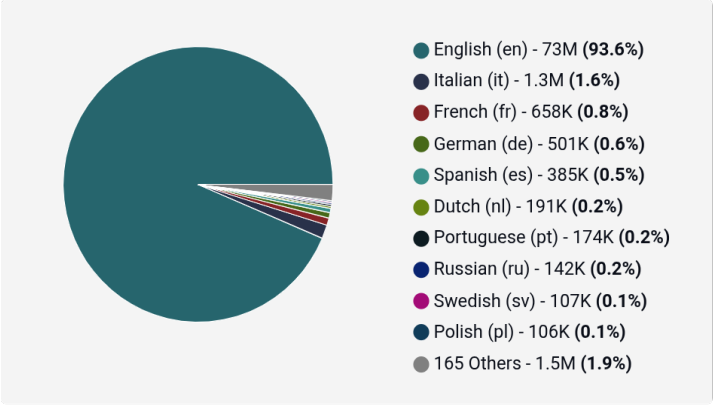
≤ 25 segments 76.58% (2.5M documents)
> 25 segments 23.42% (753K documents)

Document collections

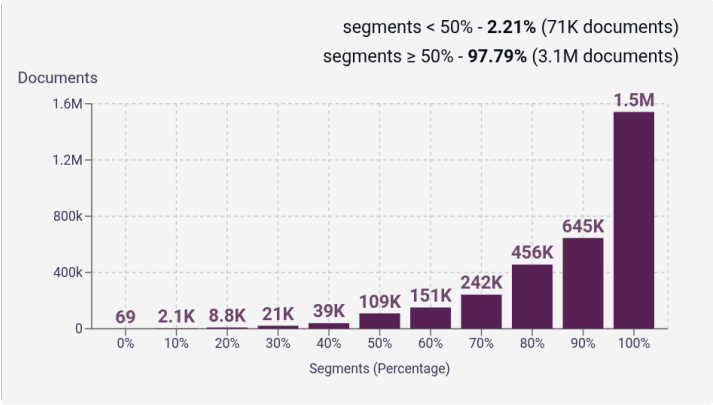


Language Distribution

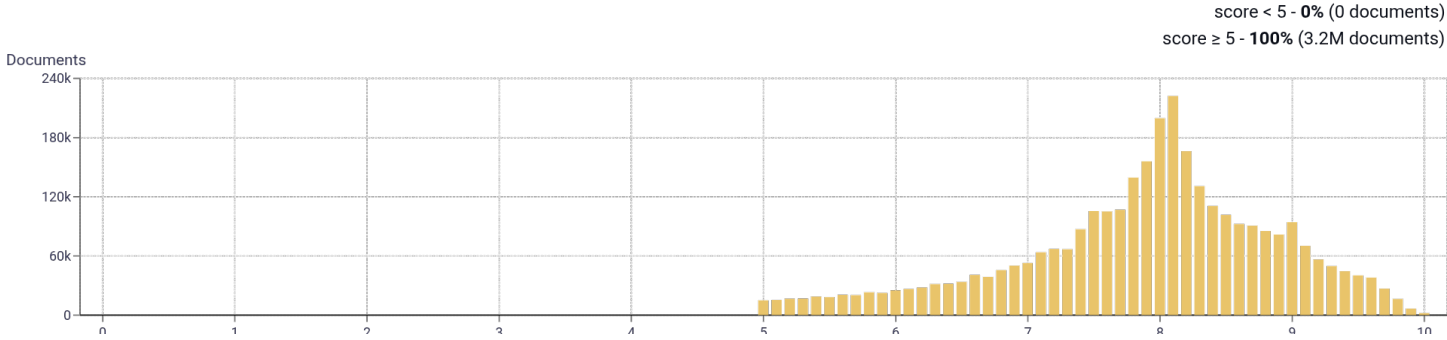
Number of segments in the English (en) corpus



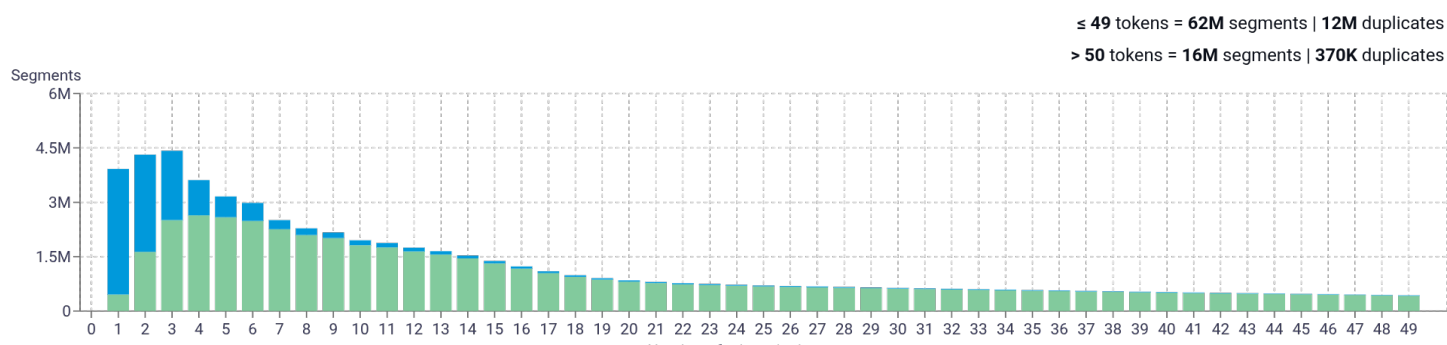
Percentage of segments in English (en) inside documents



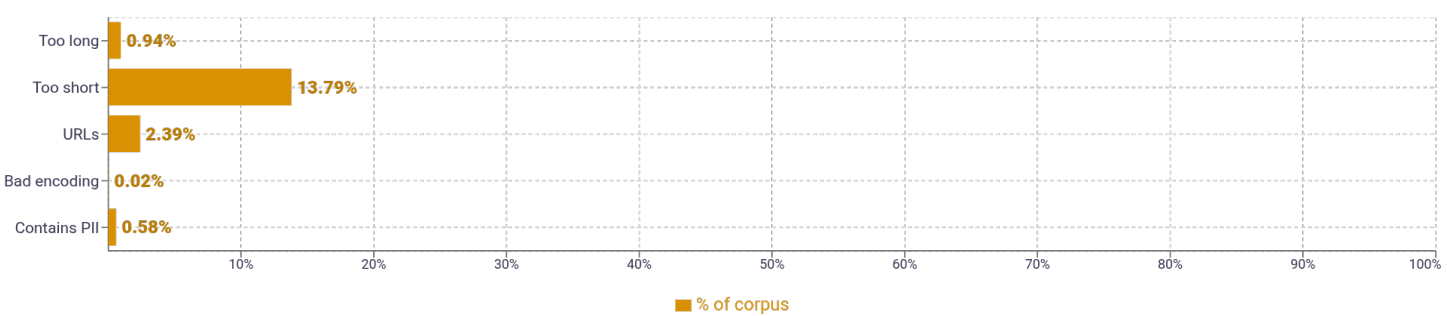
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	time 3,815,893 new 3,725,847 like 3,595,653 get 3,053,016 would 2,976,135	
2	new york 338,649 united states 332,800 make sure 301,633 years ago 202,407 social media 199,329	
3	around the world 100,436 post a comment 98,412 take a look 59,210 new york city 53,619 need to know 49,777	
4	thank you so much 30,834 end of the day 19,321 give us a call 15,316 due to the fact 13,812 thank you for sharing 13,045	
5	everything you need to know 7,595 president of the united states 6,057 keep up the good work 5,583 please feel free to contact 5,141 embodiment of the present invention 5,108	

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 *number of 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>, , , 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				