HPLT Analytics report

General overview

Corpus	Date	Language
hplt-v3-por_Latn	9/19/2025	Portuguese (pt)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
342,527,772	8,079,229,710	4,432,570,410 (54.86 %)	234B	1,235,001,507,273	1.16 TB

Top 10 domains

Domain	Docs	% of total
blogspot.com	26M	7.46%
uol.com.br	6.8M	2.00%
blogspot.com.br	5M	1.45%
wordpress.com	4.6M	1.35%
sapo.pt	4M	1.18%
globo.com	3.1M	0.90%
blogspot.pt	2.2M	0.65%
ig.com.br	1.6M	0.47%
estadao.com.br	1.2M	0.36%
docplayer.com.br	1M	0.31%

Top 10 TLDs

Domain	Docs	% of total
com.br	146M	42.58%
com	111M	32.31%
pt	27M	7.74%
net	7.5M	2.19%
org.br	7.4M	2.15%
org	7.2M	2.10%
br	4.1M	1.21%
info	1.7M	0.51%
edu.br	1.5M	0.42%
sp.gov.br	1.1M	0.32%

Register labels

- HI - 2.7%
- ID - 1.7%
- IN - 13.7%
- IP - 19.0%
- LY - 0.4%
- MIX - 4.5%
- NA - 42.8%
- OP - 7.8%
- SP - 0.4%
- UNK - 7.0%

- HI_other - 1.7%
- HI_re - 1.0%
- ID_other - 1.7%
- IN_dtp - 4.9%
- IN_en - 0.4%
- IN_fi - 0.0%
- IN_lt - 0.8%
- IN_other - 7.0%
- IN_ra - 0.5%
- IP_ds - 15.9%
- IP_other - 3.1%
- LY_other - 0.4%
- MIX - 4.5%
- NA_nb - 6.3%
- NA_ne - 29.1%
- NA_other - 3.4%
- NA_sr - 3.9%
- OP_av - 0.9%
- OP_ob - 2.1%
- OP_other - 1.7%
- OP_rs - 1.4%
- OP_rv - 1.7%
- SP_it - 0.3%
- SP_other - 0.1%
- UNK - 7.0%
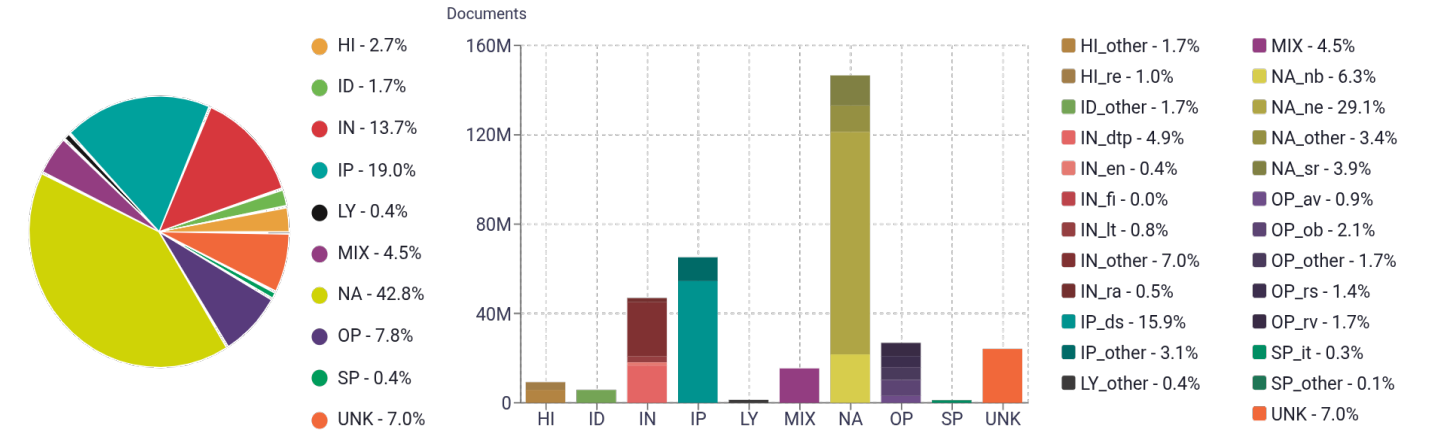
🤖 **MT**:3.3% | 11M Documents

Documents size (in segments) ⓘ

≤ 25 segments **78.35%** (268M documents)
> 25 segments **21.65%** (74M documents)

Document collections

CC = 87.34%
IA = 12.66%

67 Others (343M)
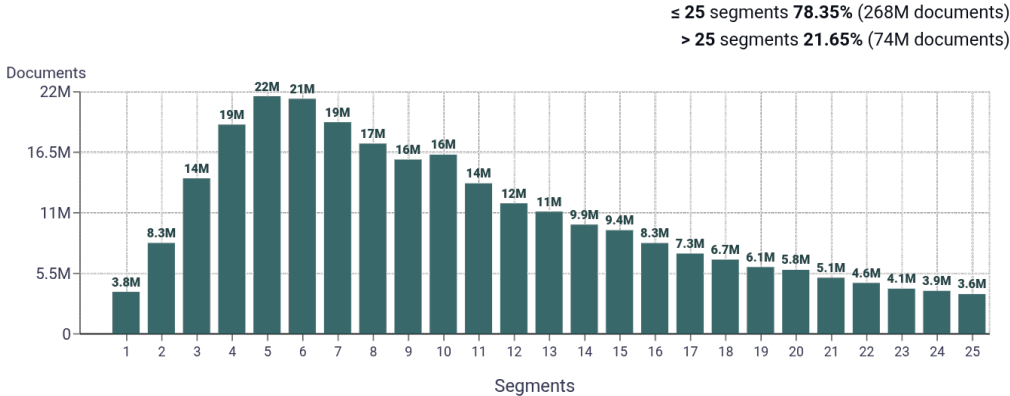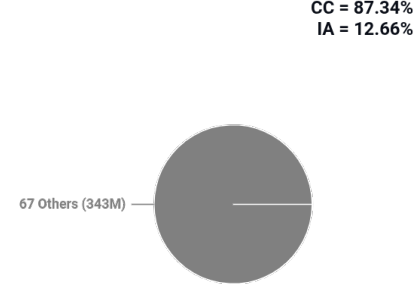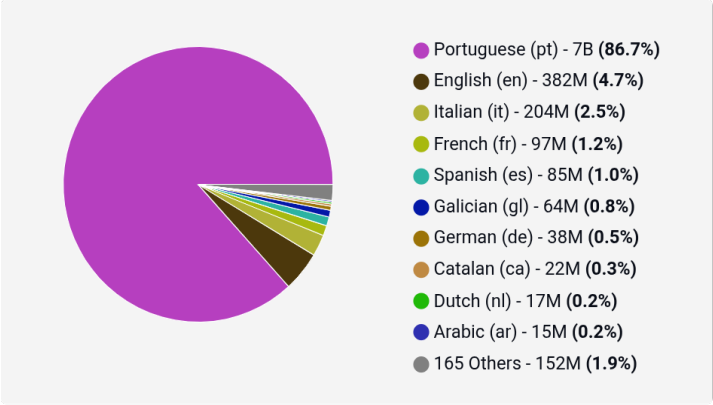
Language Distribution

Number of segments in the Portuguese (pt) corpus

- Portuguese (pt) - 7B **(86.7%)**
- English (en) - 382M **(4.7%)**
- Italian (it) - 204M **(2.5%)**
- French (fr) - 97M **(1.2%)**
- Spanish (es) - 85M **(1.0%)**
- Galician (gl) - 64M **(0.8%)**
- German (de) - 38M **(0.5%)**
- Catalan (ca) - 22M **(0.3%)**
- Dutch (nl) - 17M **(0.2%)**
- Arabic (ar) - 15M **(0.2%)**
- 165 Others - 152M **(1.9%)**

Percentage of segments in Portuguese (pt) inside documents
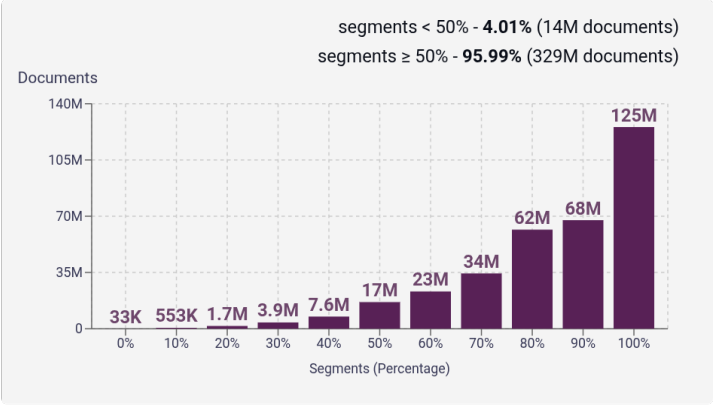
segments < 50% - **4.01%** (14M documents)
segments ≥ 50% - **95.99%** (329M documents)

Documents

Segments (Percentage)	Documents
0%	33K
10%	553K
20%	1.7M
30%	3.9M
40%	7.6M
50%	17M
60%	23M
70%	34M
80%	62M
90%	68M
100%	125M

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (343M documents)

Segment length distribution by token

≤ **49** tokens = **6.7B** segments | **3.3B** duplicates
> **50** tokens = **1.4B** segments | **355M** duplicates

Segment noise distribution

	% of corpus
Too long	0.85%
Too short	15.12%
URLs	1.84%
Bad encoding	0.00%
Contains PII	0.41%

■ % of corpus
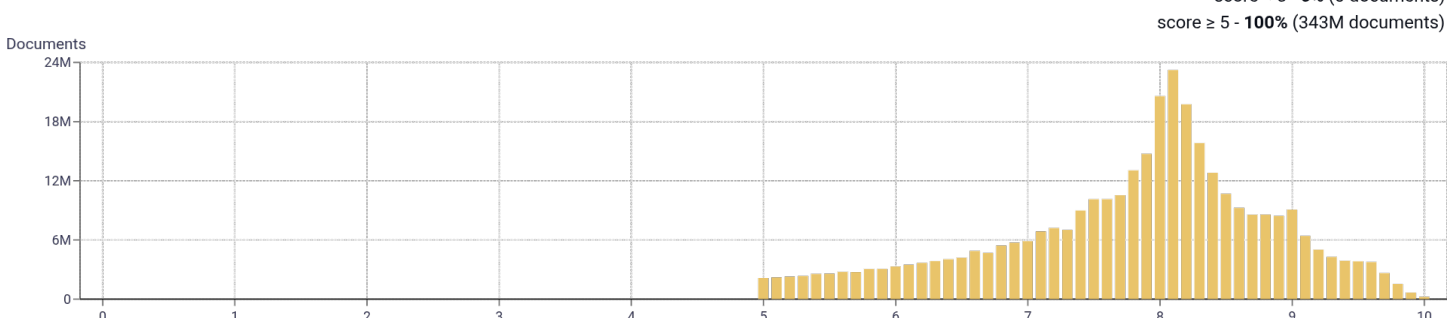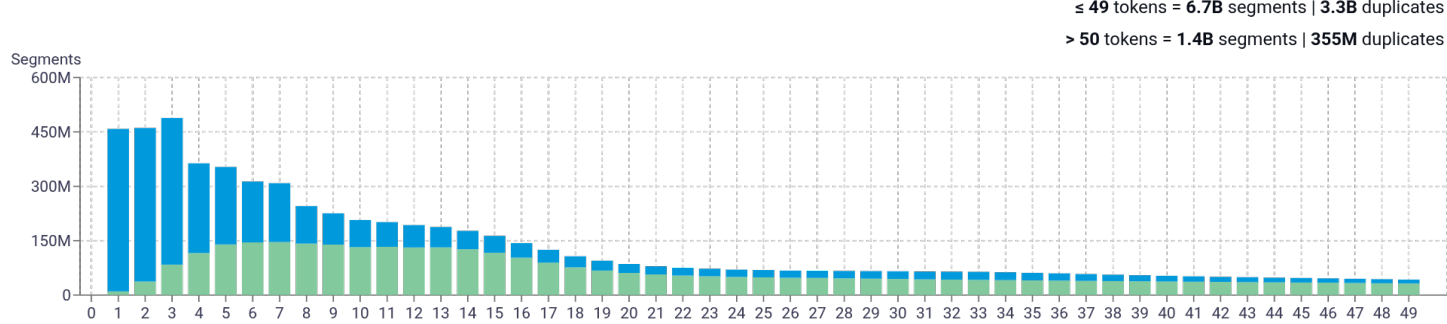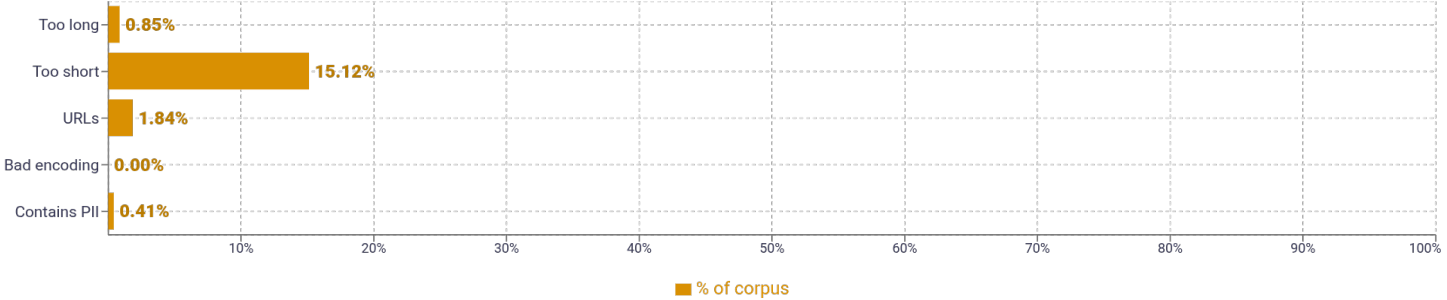
Frequent n-grams

SIZE	N-GRAMS
1	sobre | 363,466,510 pode | 348,530,544 dia | 258,778,555 ainda | 246,148,096 todos | 239,521,966
2	além disso | 49,933,641 cada vez | 28,878,720 redes sociais | 23,239,893 estados unidos | 23,055,873 muitas vezes | 19,125,524
3	rio de janeiro | 32,428,033 postar um comentário | 19,321,646 todos os dias | 11,246,571 dia a dia | 9,671,544 vale a pena | 9,361,868
4	rio grande do sul | 7,819,155 estado de são paulo | 6,107,824 rio grande do norte | 3,382,782 mato grosso do sul | 3,311,091 todos os direitos reservados | 2,474,302
5	estado do rio de janeiro | 2,265,898 luiz inácio lula da silva | 1,636,374 tudo o que você precisa | 1,442,401 trabalho de conclusão de curso | 1,292,444 melhor forma de comprar online | 1,065,702

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				