HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
gl_1.jsonl.tsv	3/21/2024	Galician (gl)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
731,356	92,682,759	19,118,996 (20.63 %)	1B	5.03 GB	

Top 10 domains

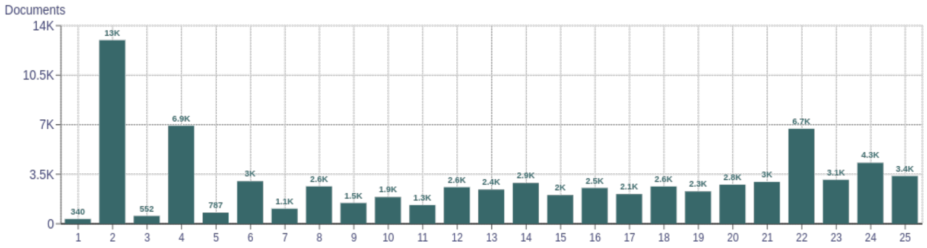
Domain	Docs	% of total
blogspot.com.es	76K	10.34
wikipedia.org	30K	4.11
blogspot.com	28K	3.79
xunta.gal	12K	1.67
lugo.gal	11K	1.49
wordpress.com	10K	1.41
pontevedraviva.com	6.5K	0.89
bretemas.gal	6.4K	0.88
vigo.org	5.8K	0.80
galipedia.com	5.4K	0.74

Top 10 TLDs

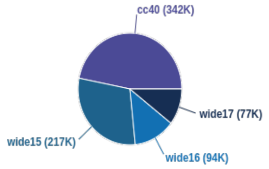
Domain	Docs	% of total
com	244K	33.34
gal	138K	18.85
org	115K	15.77
es	98K	13.36
com.es	76K	10.36
info	10K	1.38
eu	7.8K	1.06
net	7.6K	1.04
com.ar	5.9K	0.81
pt	3.1K	0.42

Documents size (in segments)

<= 25 segments **10.36%** (76K documents)
> 25 segments **89.64%** (655K documents)

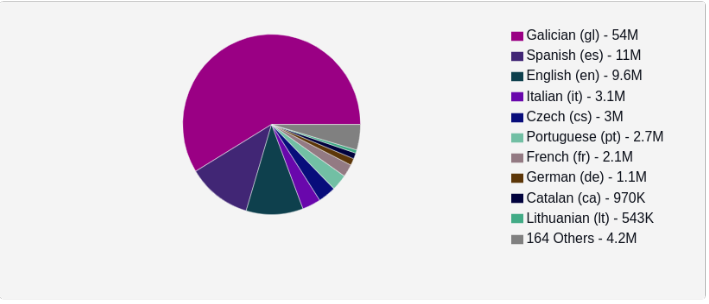


Documents by collection

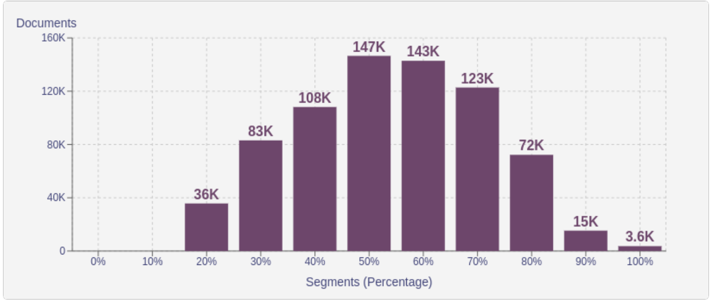


Language Distribution

Number of segments

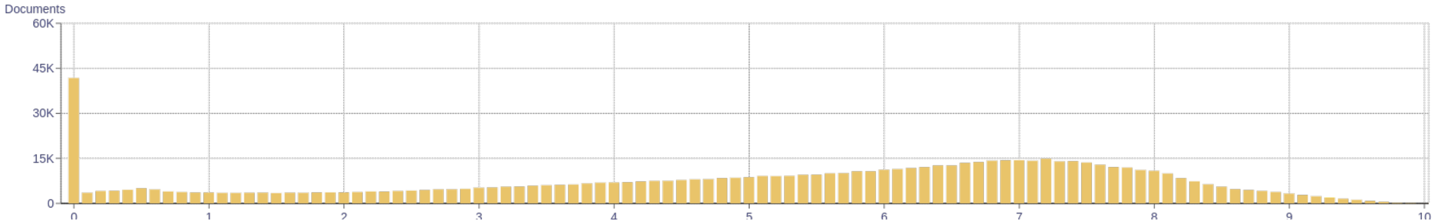


Percentage of segments in Galician (gl) inside documents



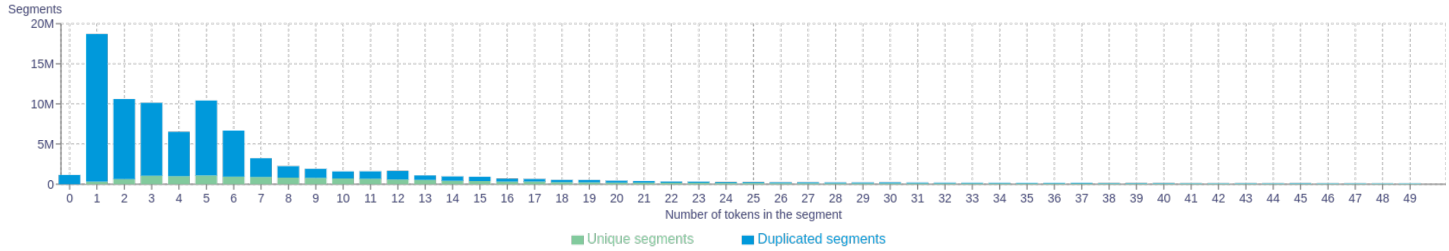
Distribution of documents by document score

score <= 5 - **40.14%** (294K documents)
score > 5 - **59.86%** (438K documents)

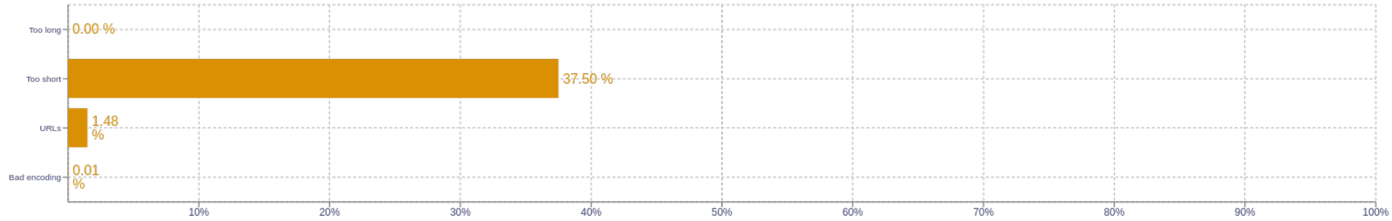


Segment length distribution by token

<= 49 tokens = **16M** segments | **72M** duplicates
> 50 tokens = **4.3M** segments | **1.2M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	y | 2371096 galicia | 1898093 día | 1110181 abril | 1085104 marzo | 1051215
2	hay comentarios | 300026 correo electrónicoescribe | 292932 correo electrónico | 231870 aviso legal | 220674 sitio web | 215834
3	enviar por correo | 450385 facebookcompartir en pinterest | 433535 blogcompartir con twittercompartir | 292942 electrónicoescribe un blogcompartir | 292931 twittercompartir con facebookcompartir | 289851
4	enviar por correo electrónicoescribe | 292932 correo electrónicoescribe un blogcompartir | 292931 enviar por correo electrónicoblogthis | 143634 enlaces a esta entrada | 68538 entradas antiguas página principal | 53522
5	electrónicoescribe un blogcompartir con twittercompartir | 292931 blogcompartir con twittercompartir con facebookcompartir | 289851 twittercompartir con facebookcompartir en pinterest | 289850 compartir en twittercompartir en facebookcompartir | 143696 twittercompartir en facebookcompartir en pinterest | 143685

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