HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
glg_Latn.jsonl.tsv	9/21/2024	Galician (gl)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
3,020,164	61,177,888	25,143,278 (41.10 %)	1.9B	9.6 GB	10,050,502,462

Top 10 domains

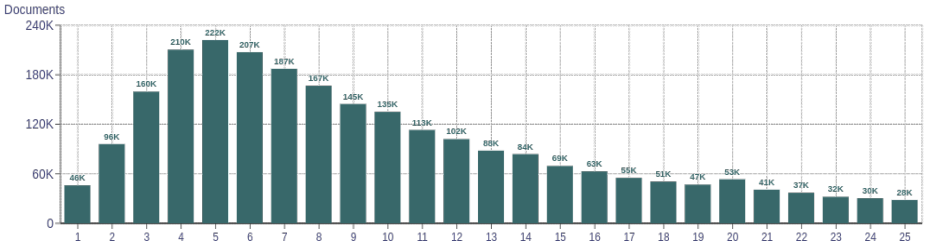
Domain	Docs	% of total
wikipedia.org	379K	12.56
blogspot.com	240K	7.96
blogspot.com.es	115K	3.81
wordpress.com	85K	2.81
xunta.gal	53K	1.76
crtvg.es	35K	1.15
bng.gal	31K	1.02
pontevedraviva.com	28K	0.93
blogaliza.org	24K	0.78
vieiros.com	21K	0.68

Top 10 TLDs

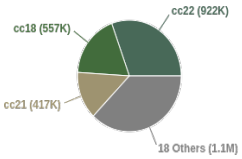
Domain	Docs	% of total
com	1.1M	35.72
org	693K	22.95
gal	561K	18.57
es	380K	12.60
com.es	115K	3.82
net	42K	1.39
eu	35K	1.16
info	26K	0.85
gl	9.7K	0.32
com.ar	9K	0.30

Documents size (in segments)

<= 25 segments **81.75%** (2.5M documents)
> 25 segments **18.25%** (551K documents)

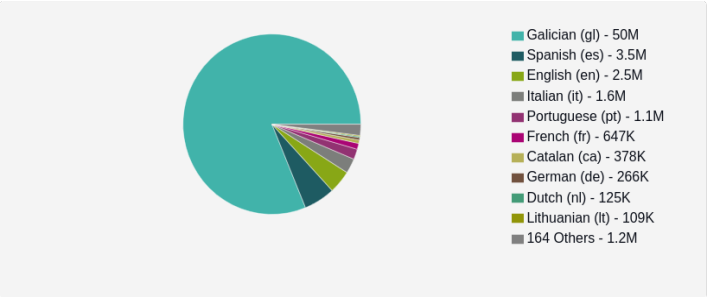


Documents by collection

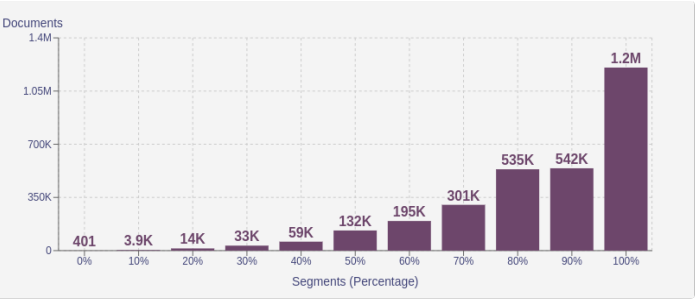


Language Distribution

Number of segments

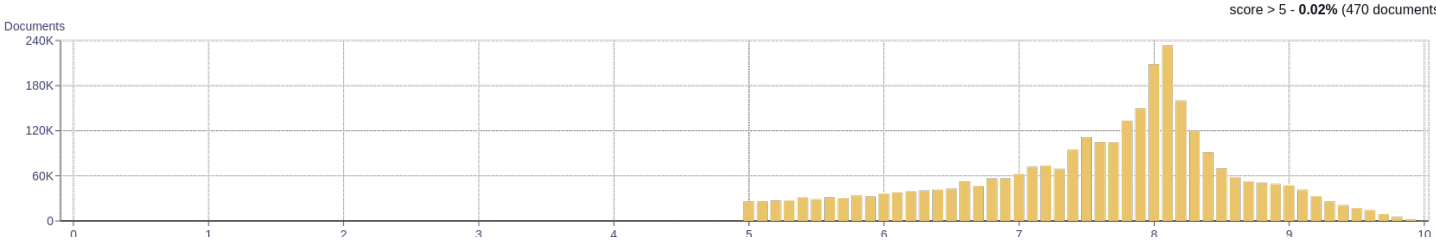


Percentage of segments in Galician (gl) inside documents



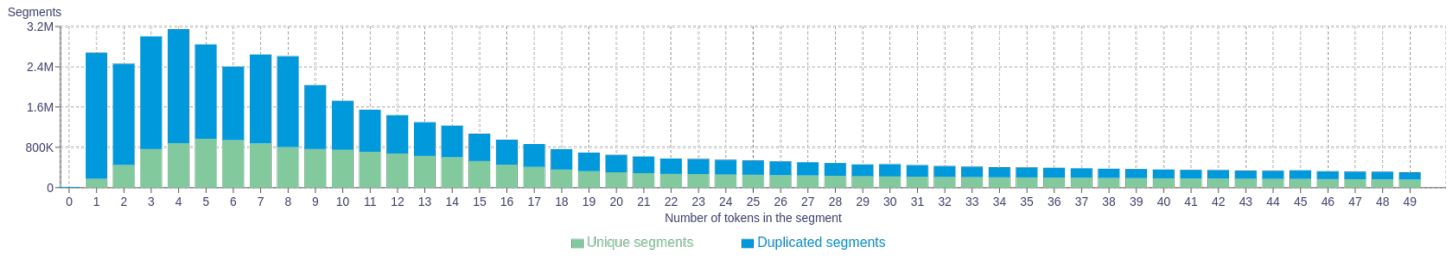
Distribution of documents by document score

score <= 5 - **99.98%** (3M documents)
score > 5 - **0.02%** (470 documents)

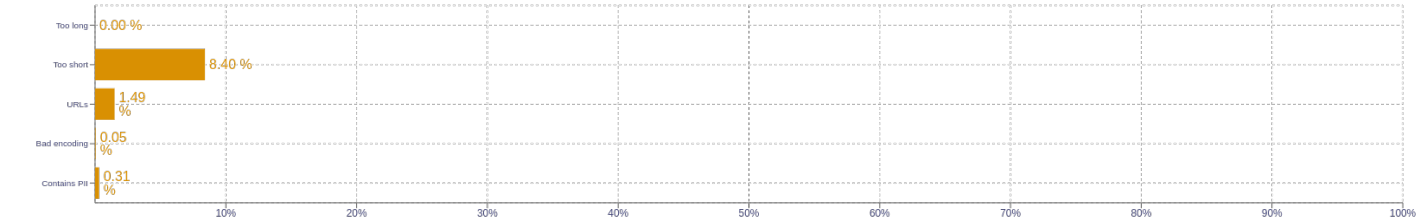


Segment length distribution by token

<= 49 tokens = **18M** segments | **30M** duplicates
> 50 tokens = **13M** segments | **6.2M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	editar | 3177972 · entre | 2709020 · galicia | 2364440 · anos | 2200798 · ano | 1929429
2	estados unidos | 210752 · medio ambiente | 155997 · primeira vez | 142910 · terá lugar | 138886 · lingua galega | 127960
3	editar a fonte | 1491606 · santiago de compostela | 431155 · xunta de galicia | 278707 · millóns de euros | 160238 · fin de semana | 139122
4	arquivado dende o orixinal | 63686 · comunidade autónoma de galicia | 48245 · día das letras galegas | 44579 · diario oficial de galicia | 39354 · consello da cultura galega | 31830
5	universidade de santiago de compostela | 55810 · contra a violencia de xénero | 21673 · prazo de presentación de solicitudes | 20239 · electrónica da xunta de galicia | 19123 · dende o punto de vista | 15602

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