

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

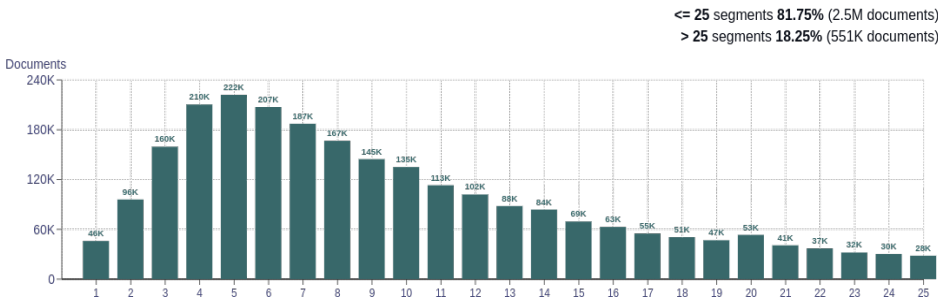
Top 10 domains

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

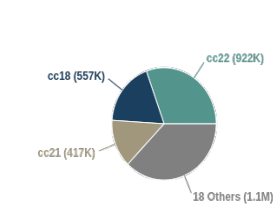
Top 10 TLDs

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)

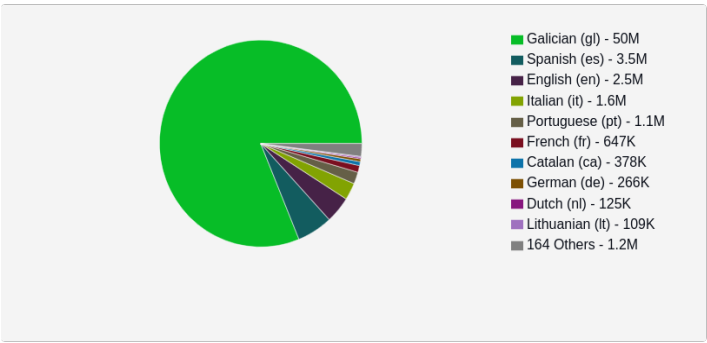


Documents by collection

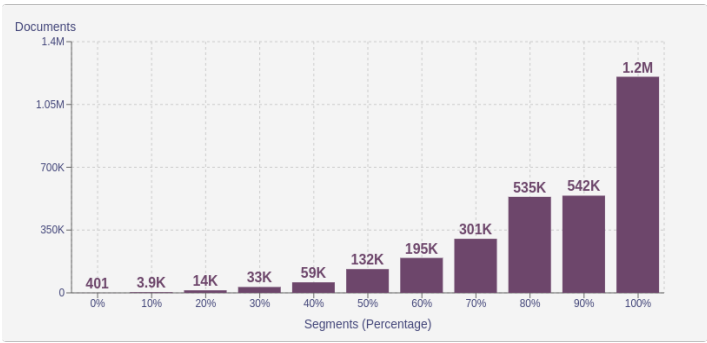


Language Distribution

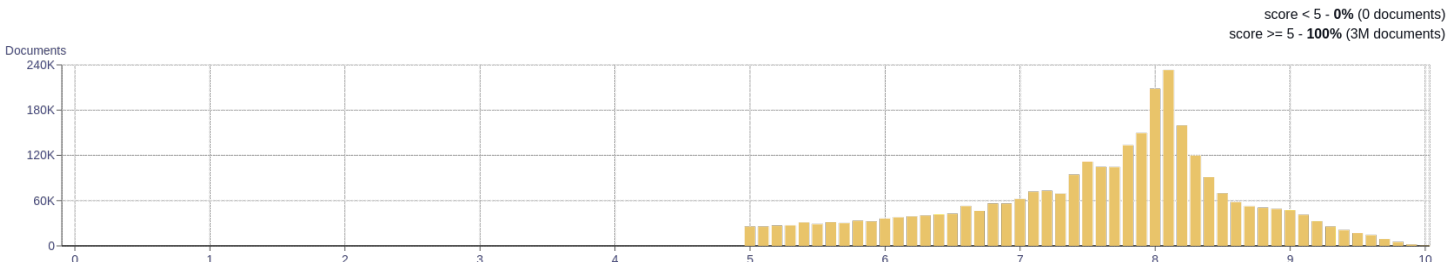
Number of segments



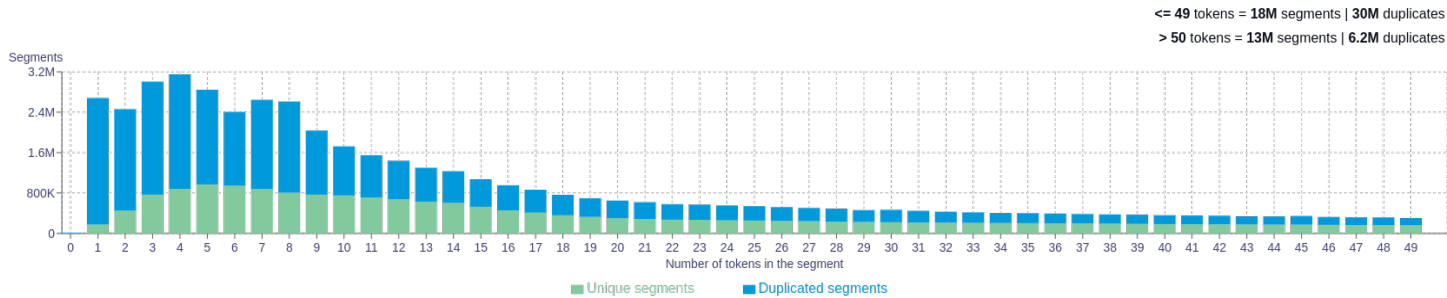
Percentage of segments in Galician (gl) inside documents



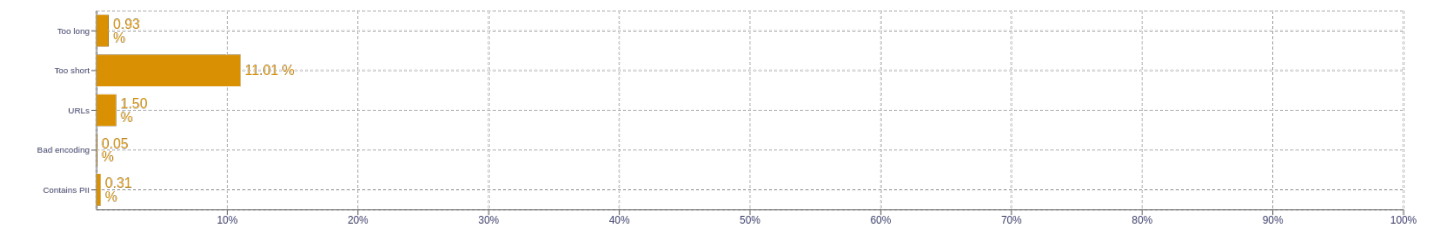
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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