HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
eu_1.jsonl.tsv	3/20/2024	Basque (eu)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
343,947	37,226,281	10,165,741 (27.31 %)	412M	2.3 GB	

Top 10 domains

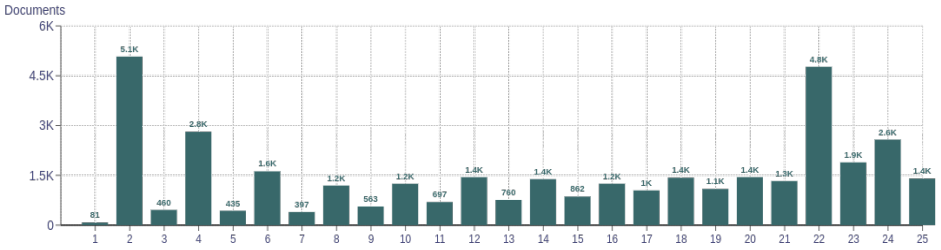
Domain	Docs	% of total
blogspot.com.es	25K	7.22
wikipedia.org	13K	3.83
hitza.eus	6.4K	1.87
blogspot.com	6.2K	1.80
lab.eus	6K	1.75
argia.eus	5.7K	1.66
berria.eus	4.4K	1.27
dantzan.eus	4.4K	1.27
euskadi.eus	3.9K	1.13
zuzeu.eus	3.2K	0.92

Top 10 TLDs

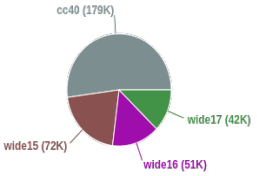
Domain	Docs	% of total
eus	146K	42.50
com	78K	22.56
org	42K	12.34
com.es	25K	7.25
es	17K	4.91
net	12K	3.63
info	5.3K	1.54
eu	2.3K	0.68
biz	1.9K	0.55
io	1.6K	0.48

Documents size (in segments)

<= 25 segments **10.84%** (37K documents)
> 25 segments **89.16%** (307K documents)

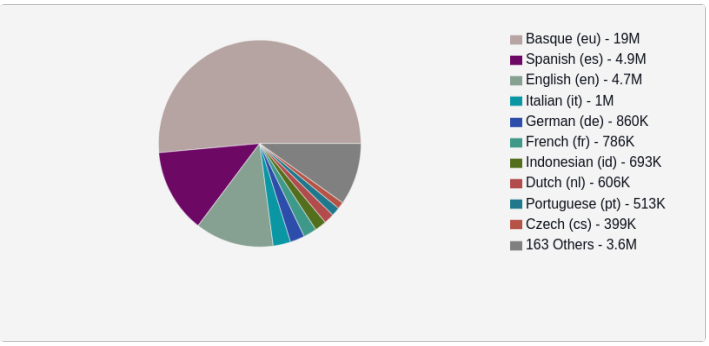


Documents by collection

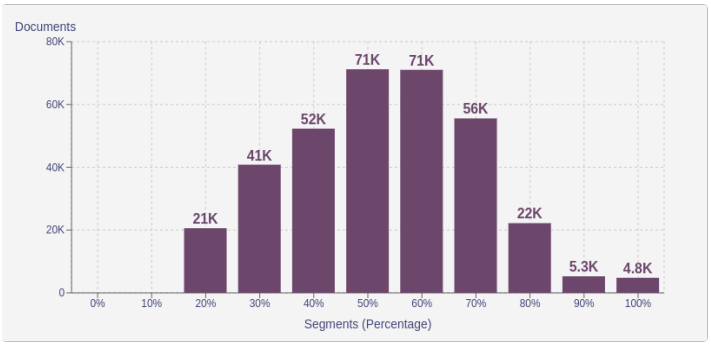


Language Distribution

Number of segments

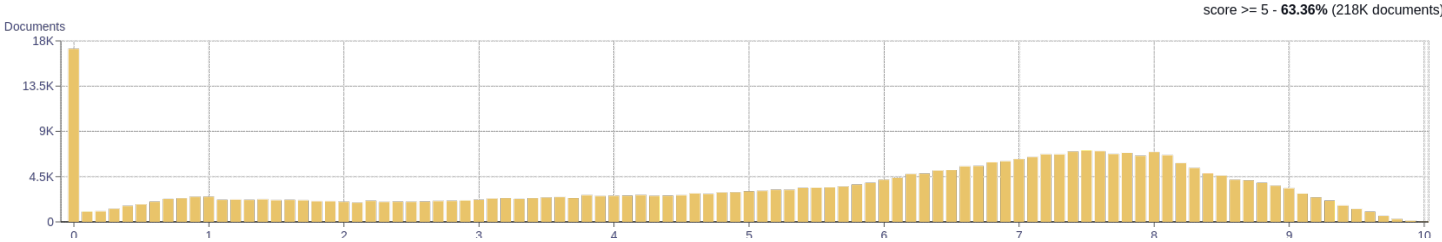


Percentage of segments in Basque (eu) inside documents



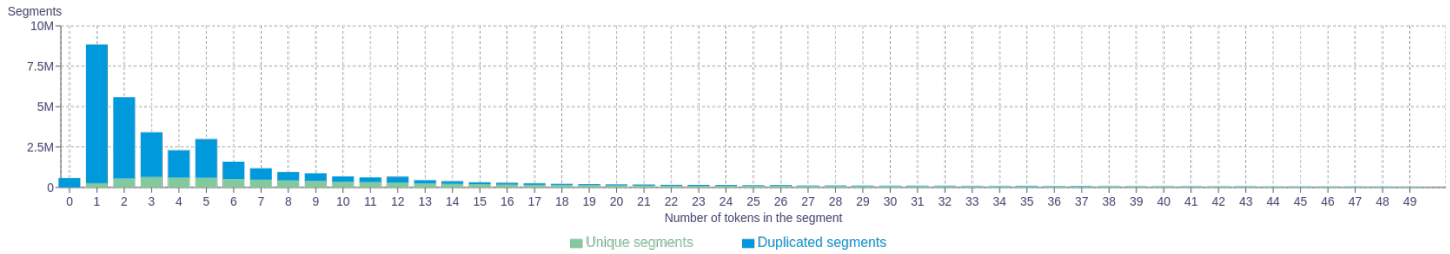
Distribution of documents by document score

score < 5 - **36.64%** (126K documents)
score >= 5 - **63.36%** (218K documents)

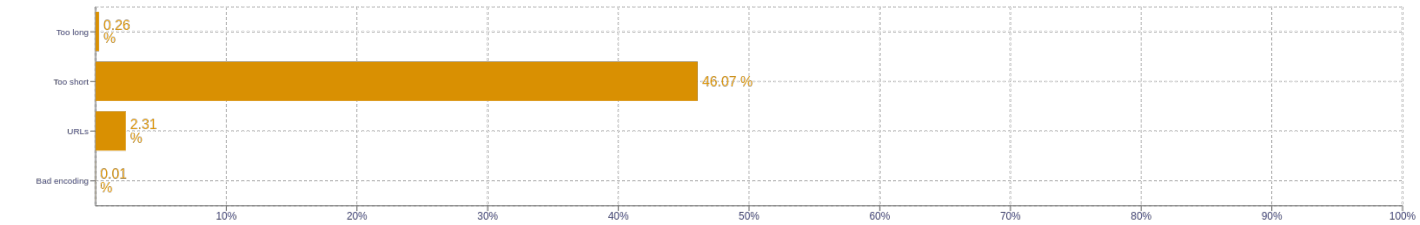


Segment length distribution by token

<= 49 tokens = **8.8M** segments | **27M** duplicates
> 50 tokens = **1.8M** segments | **416K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	de | 4070563 · la | 1959228 · en | 1634840 · y | 1349054 · el | 1304858
2	de la | 497473 · en el | 218700 · en la | 189454 · de los | 165211 · a la | 148796
3	no hay comentarios | 93711 · enviar por correo | 87288 · con twittercompartir con | 86480 · un blogcompartir con | 86479 · blogcompartir con twittercompartir | 86479
4	un blogcompartir con twittercompartir | 86479 · blogcompartir con twittercompartir con | 86479 · por correo electrónicoescribe un | 86472 · enviar por correo electrónicoescribe | 86472 · electrónicoescribe un blogcompartir con | 86472
5	un blogcompartir con twittercompartir con | 86479 · por correo electrónicoescribe un blogcompartir | 86472 · enviar por correo electrónicoescribe un | 86472 · electrónicoescribe un blogcompartir con twittercompartir | 86472 · correo electrónicoescribe un blogcompartir con | 86472

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