HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
eus_Latn.jsonl.tsv	9/6/2024	Basque (eu)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,974,218	37,621,611	17,454,038 (46.39 %)	949M	5.63 GB	6,016,518,017

Top 10 domains

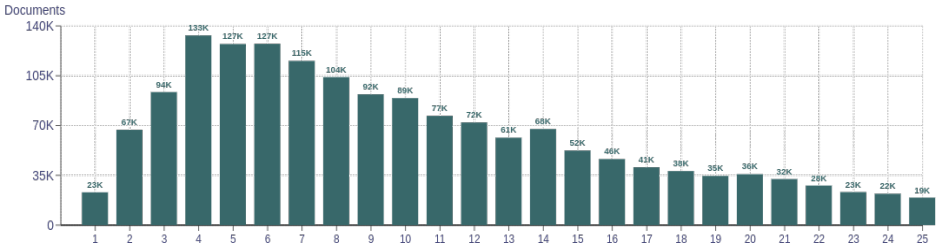
Domain	Docs	% of total
wikipedia.org	322K	16.30
argia.eus	51K	2.57
blogspot.com	40K	2.00
zuzeu.eus	31K	1.58
berria.eus	29K	1.48
euskadi.eus	28K	1.40
blogspot.com.es	27K	1.37
hitza.info	26K	1.30
eitb.eus	23K	1.19
consumer.es	22K	1.10

Top 10 TLDs

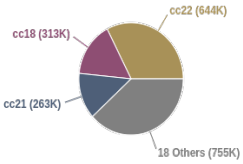
Domain	Docs	% of total
eus	812K	41.15
org	452K	22.91
com	400K	20.28
es	96K	4.87
net	74K	3.73
info	49K	2.48
com.es	27K	1.39
eu	11K	0.58
fr	7K	0.36
biz	6.8K	0.34

Documents size (in segments)

<= 25 segments **82.24%** (1.6M documents)
> 25 segments **17.76%** (351K documents)

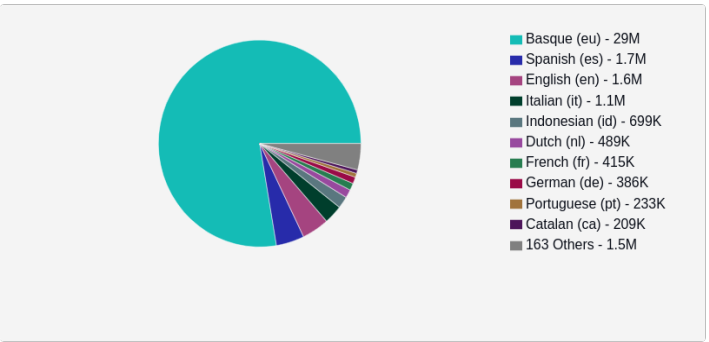


Documents by collection

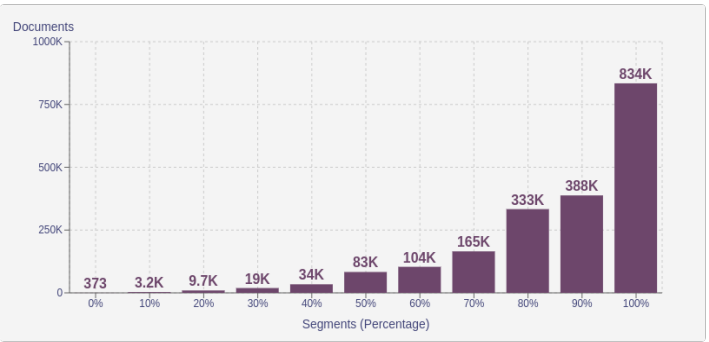


Language Distribution

Number of segments

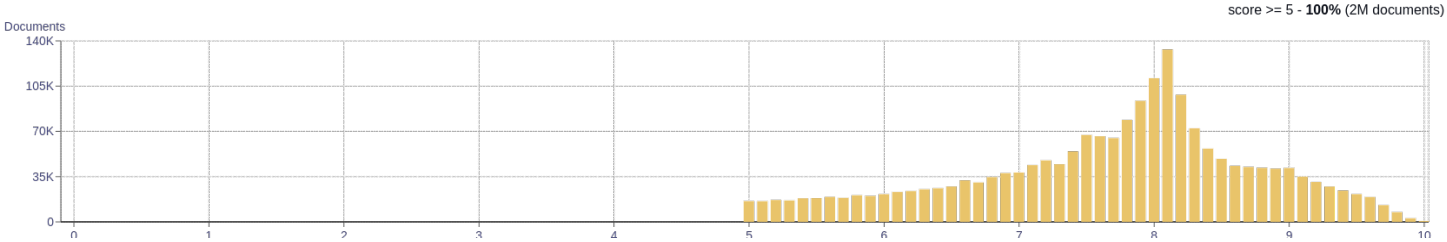


Percentage of segments in Basque (eu) inside documents



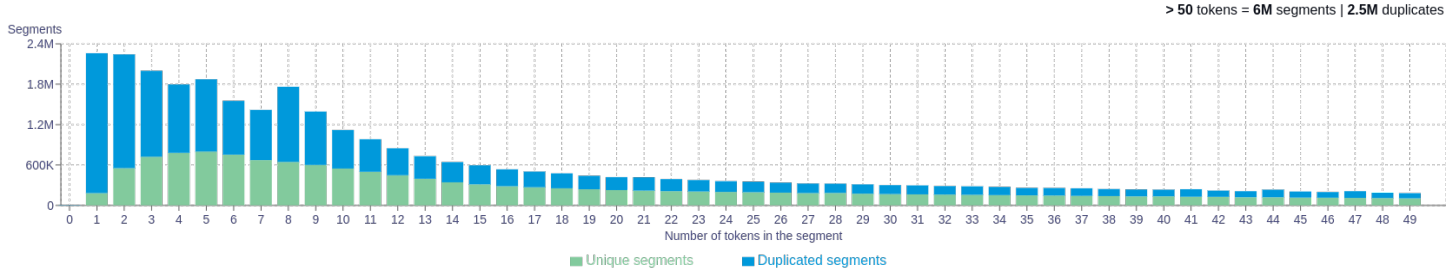
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (2M documents)

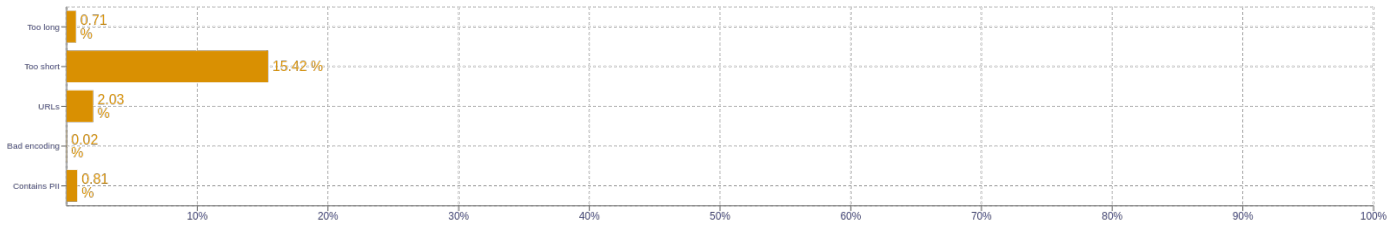


Segment length distribution by token

<= 49 tokens = **14M** segments | **18M** duplicates
> 50 tokens = **6M** segments | **2.5M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	de | 2432342 · behar | 2341765 · aldatu | 2157354 · izango | 1971789 · egiten | 1890172
2	iturburu kodea | 922361 · aldatu iturburu | 921849 · de la | 311508 · euskal herriko | 259874 · ahal izango | 248958
3	aldatu iturburu kodea | 921847 · artikulu honen edukiaren | 68551 · zati bat lur | 68151 · lur entziklopedia tematikotik | 68145 · lur hiztegi entziklopedikotik | 68136
4	artikulu honen edukiaren zati | 68550 · edukiaren zati bat lur | 68137 · entziklopedikotik edo lur entziklopedia | 68135 · hiztegi entziklopedikotik edo lur | 68134 · zati bat lur hiztegi | 68133
5	entziklopedikotik edo lur entziklopedia tematikotik | 68135 · lur hiztegi entziklopedikotik edo lur | 68134 · hiztegi entziklopedikotik edo lur entziklopedia | 68134 · zati bat lur hiztegi entziklopedikotik | 68133 · edukiaren zati bat lur hiztegi | 68133

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