HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
HPLT-v2-cat_Latn.tsv	9/22/2024	Catalan (ca)

Volumes

Docs	Segments	Characters	Size
18,553,883	383,333,998	59,821,587,389	57.38 GB

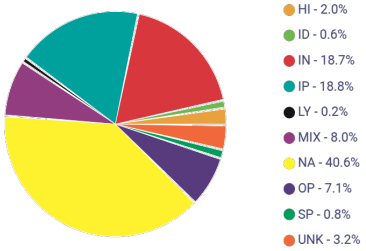
Top 10 domains

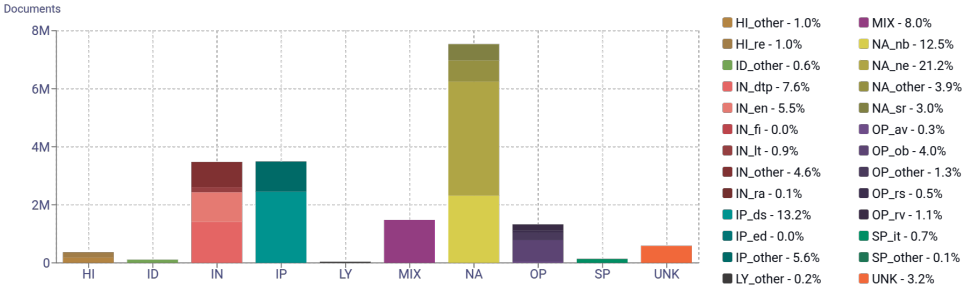
Domain	Docs	% of total
blogspot.com	1.7M	8.96%
wikipedia.org	936K	5.04%
blogspot.com.es	664K	3.58%
wordpress.com	447K	2.41%
ara.cat	217K	1.17%
ccma.cat	142K	0.77%
gencat.cat	122K	0.66%
diaridegirona.cat	120K	0.64%
regio7.cat	104K	0.56%
agoda.com	92K	0.50%

Top 10 TLDs

Domain	Docs	% of total
cat	6.9M	37.21%
com	6.4M	34.29%
org	2.1M	11.56%
es	899K	4.84%
com.es	666K	3.59%
net	498K	2.69%
edu	209K	1.13%
info	153K	0.82%
ad	141K	0.76%
eu	71K	0.38%

Register labels

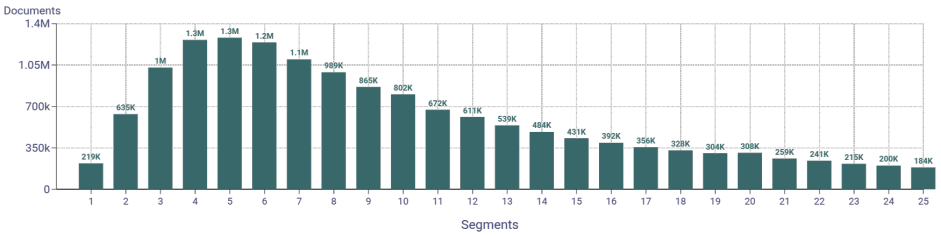




MT:1.0% | 178K Documents

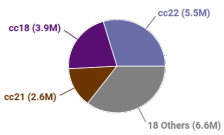
Documents size (in segments)

<= 25 segments **80.54%** (15M documents)
> 25 segments **19.46%** (3.6M documents)



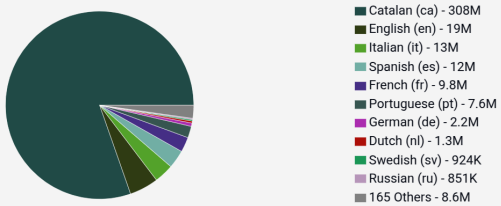
Documents by collection

CC = 74.56%
IA = 25.44%

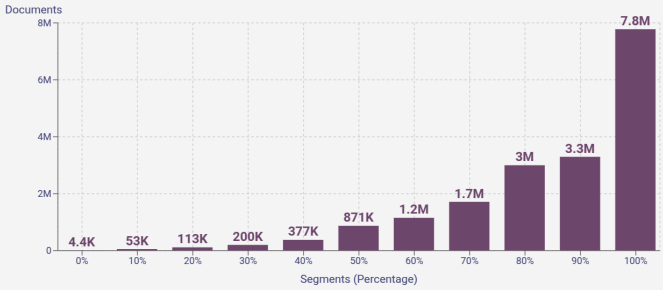


Language Distribution

Number of segments in the Catalan (ca) corpus

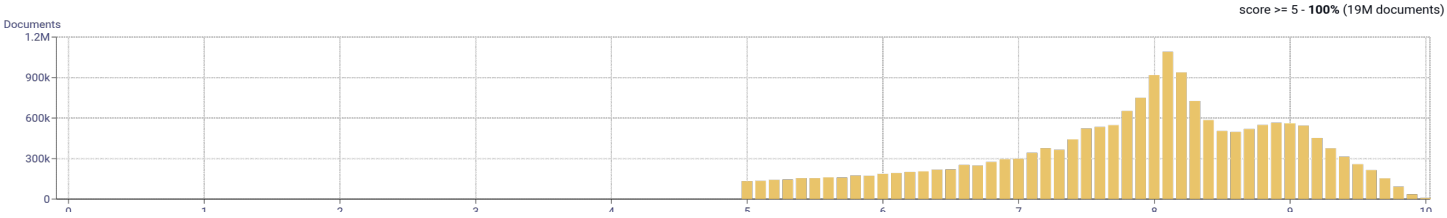


Percentage of segments in Catalan (ca) inside documents

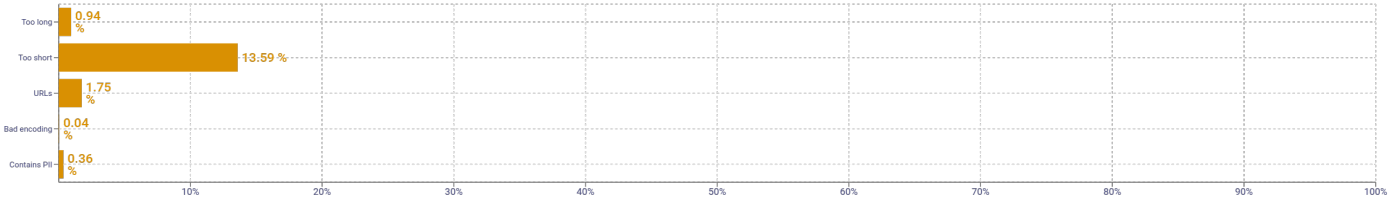


Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (19M documents)



Segment noise distribution



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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av