HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
pap_Latn.jsonl.tsv	12/3/2024	Papiamento (pap)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
89,812	1,387,382	814,803 (58.73 %)	53M	252,796,043	244.39 MB

Top 10 domains

Domain	Docs	% of total
diario.aw	4.8K	5.30
kikotapasando.com	4.3K	4.74
wikipedia.org	4.2K	4.67
masnoticia.com	4.2K	4.66
arubanative.com	3.8K	4.26
live99fm.com	3.4K	3.75
awe24.com	3.3K	3.68
noticiacla.com	2.6K	2.94
awemainta.com	2.2K	2.41
1noticia.com	1.6K	1.83

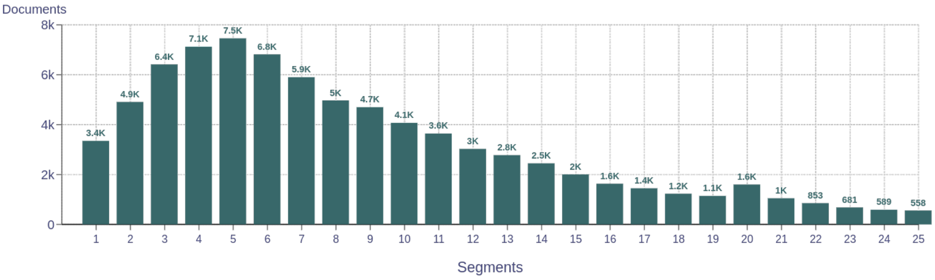
Top 10 TLDs

Domain	Docs	% of total
com	57K	63.50
aw	10K	11.08
org	9.4K	10.43
cw	3.8K	4.22
nl	3.4K	3.81
net	1.6K	1.78
nu	1.4K	1.51
news	1.2K	1.31
today	276	0.31
blog	230	0.26

Documents size (in segments)

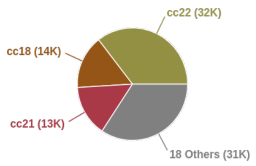
<= 25 segments **89.54%** (80K documents)

> 25 segments **10.46%** (9.4K documents)



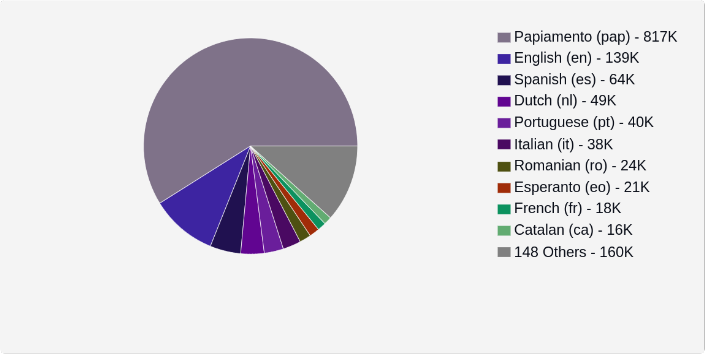
Documents by collection

CC = 72.97%
IA = 27.03%

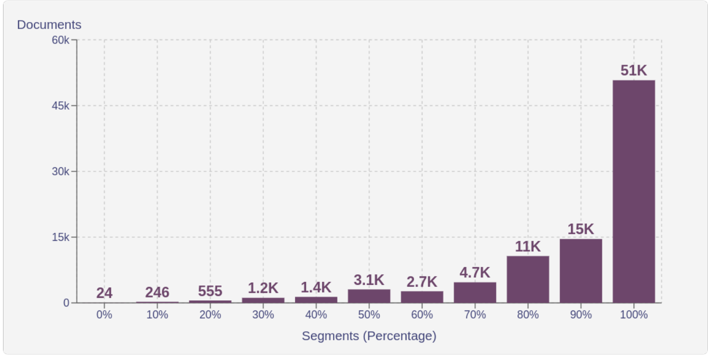


Language Distribution

Number of segments in the Papiamento (pap) corpus



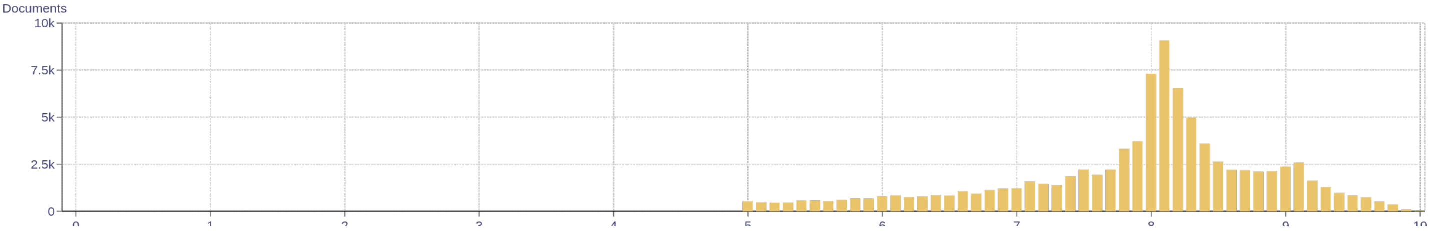
Percentage of segments in Papiamento (pap) inside documents



Distribution of documents by document score

score < 5 - **0%** (0 documents)

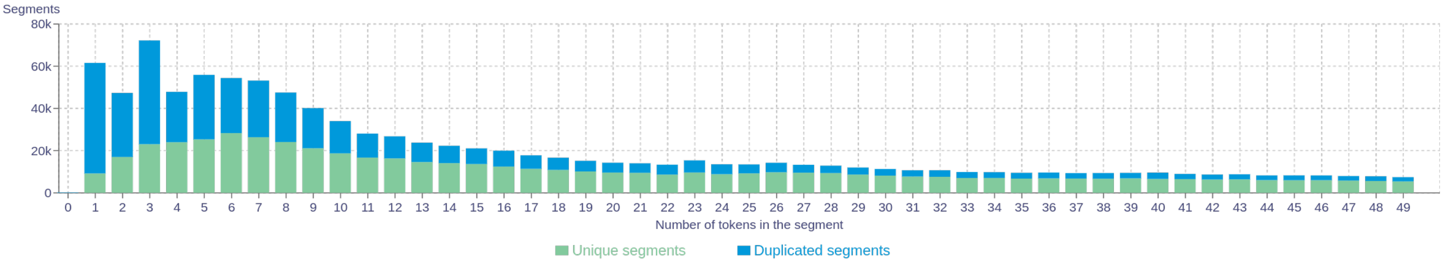
score >= 5 - **100%** (90K documents)



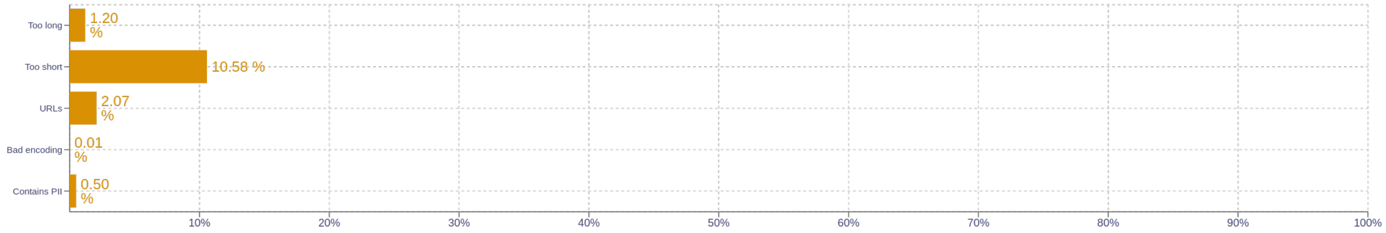
Segment length distribution by token

≤ 49 tokens = **562K** segments | **463K** duplicates

> 50 tokens = **362K** segments | **109K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	pa | 1117468 na | 749639 y | 733878 i | 474986 nan | 381363
2	na aruba | 35903 pa asina | 23408 pa nan | 21416 su mes | 19814 tur hende | 16545
3	na e sitio | 7627 el a bisa | 6354 gobierno di aruba | 5318 loke ta trata | 5225 na un manera | 4495
4	pa loke ta trata | 5009 prueba prueba prueba prueba | 2697 na e momentonan aki | 2015 prome minister evelyn wever | 2009 yegada di e patruya | 1518
5	prueba prueba prueba prueba prueba | 2559 na yegada di e patruya | 1509 camara di comercio y industria | 1227 impuesto di vehiculo di motor | 944 comercio y industria di aruba | 881

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