HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
hat_Latn.jsonl.tsv	9/27/2024	Haitian Creole (ht)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
212,686	4,635,460	2,539,373 (54.78 %)	143M	623.02 MB	634,488,750

Top 10 domains

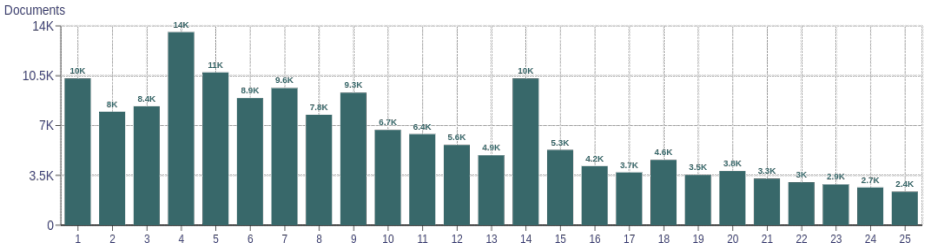
Domain	Docs	% of total
voanouvel.com	8.5K	4.00
itsmygame.org	6.5K	3.06
wikipedia.org	5.5K	2.58
temoignages.re	4.7K	2.19
studybible.info	3.7K	1.75
fouyebible.com	3.5K	1.62
makeoverarcade.com	3.4K	1.59
wondershare.com	3.4K	1.58
jw.org	3.3K	1.55
vessoft.com	3K	1.40

Top 10 TLDs

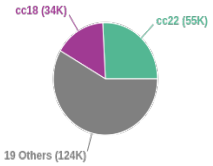
Domain	Docs	% of total
com	130K	61.32
org	37K	17.41
net	9.2K	4.35
info	7K	3.29
re	4.7K	2.22
news	3.4K	1.60
gov	3.1K	1.47
ru	1.3K	0.62
ws	984	0.46
zone	984	0.46

Documents size (in segments)

<= 25 segments **75.11%** (160K documents)
> 25 segments **24.89%** (53K documents)

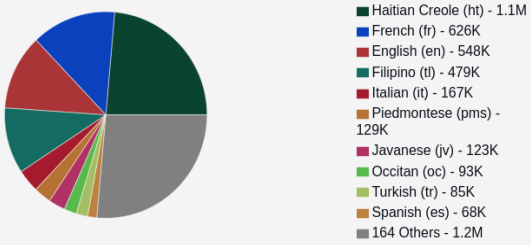


Documents by collection

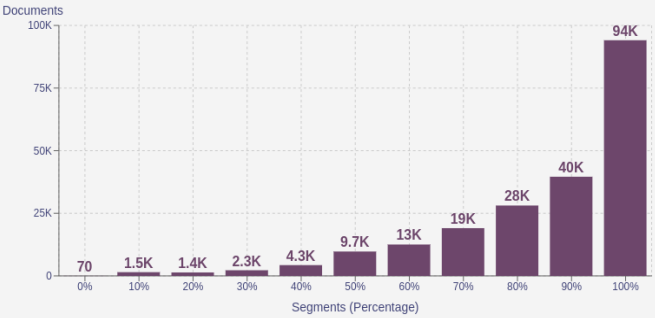


Language Distribution

Number of segments

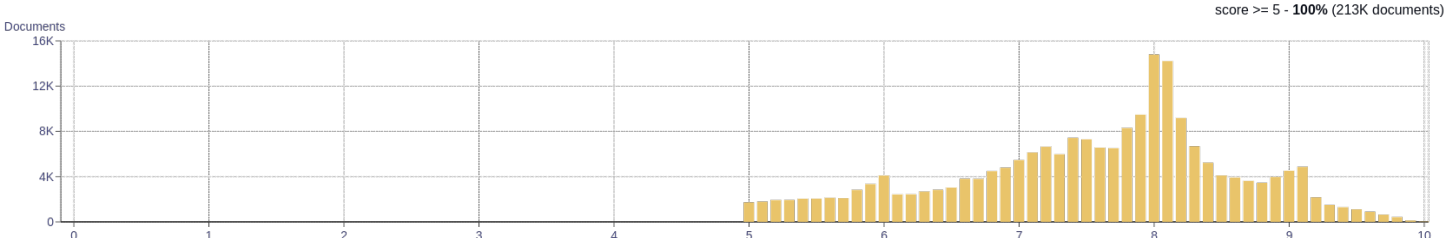


Percentage of segments in Haitian Creole (ht) inside documents



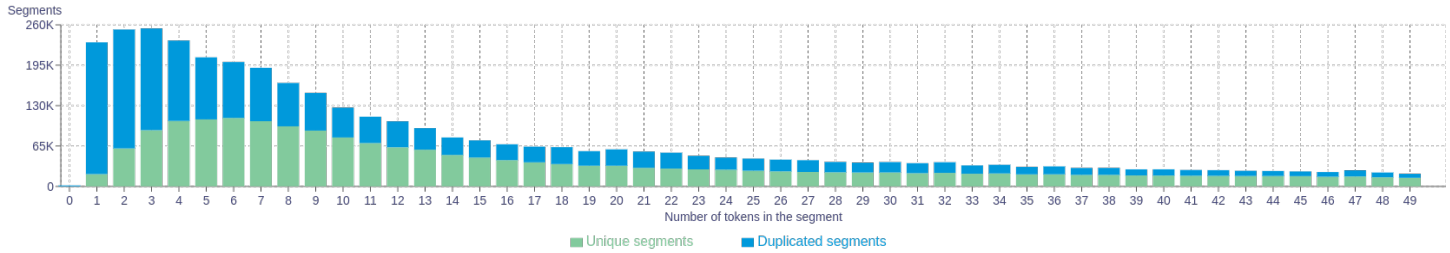
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (213K documents)

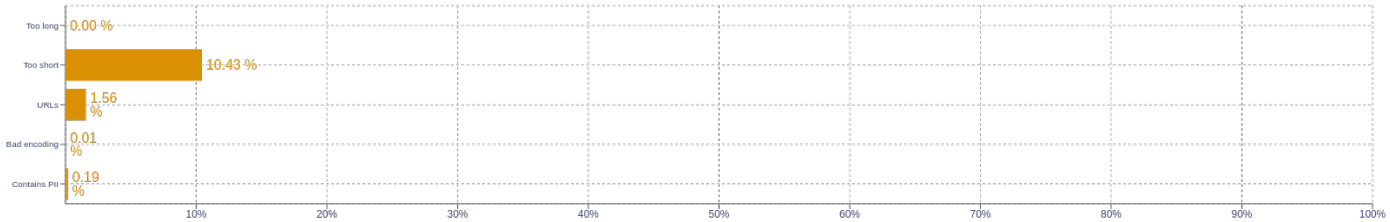


Segment length distribution by token

<= 49 tokens = **1.9M** segments | **1.9M** duplicates
> 50 tokens = **847K** segments | **240K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	moun | 713554 · jwèt | 441962 · ti | 379440 · bon | 314907 · bay | 253794
2	sit entènèt | 47513 · jwe jwèt | 26454 · ti fi | 25170 · mr speaker | 22265 · pitit gason | 22004
3	bon jan kalite | 27373 · jwèt sou entènèt | 12686 · fin vye granmoun | 10219 · jwe sou entènèt | 9819 · lojisyèl an pèmèt | 9081
4	pwopòsyon ki pi wo | 7456 · popilasyon an pi wo | 7278 · jwe sou entènèt flash | 6184 · jwèt la te jwe | 5652 · moun ki nan jwèt | 5599
5	jwe sou entènèt flash jwèt | 6180 · lyen ki nan yon zanmi | 5585 · pataje jwèt la ak mond | 5584 · kòd la html nan sit | 5584 · kopi kòd la ak keratin | 5584

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