HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
tl_1.jsonl.tsv	3/24/2024	Filipino (tl)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
585,237	104,222,137	24,808,129 (23.80 %)	1.1B	5.11 GB	

Top 10 domains

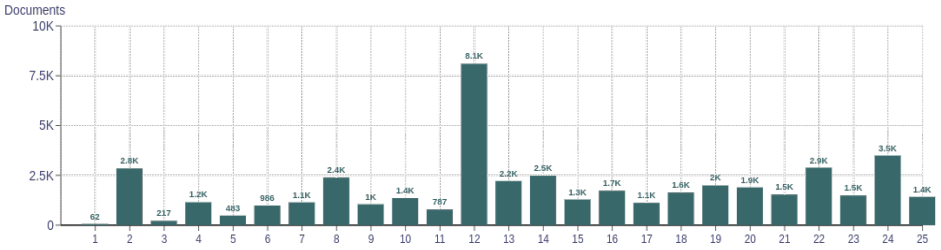
Domain	Docs	% of total
zipcodecountry.com	48K	8.19
fanpop.com	19K	3.29
depinisyon.com	15K	2.60
blogspot.com	15K	2.59
booking.com	10K	1.73
androidappsgame.net	9.9K	1.68
abs-cbn.com	8.6K	1.47
inquirer.net	8.5K	1.46
dwiz882am.com	8.5K	1.44
airbnb.com	7.8K	1.33

Top 10 TLDs

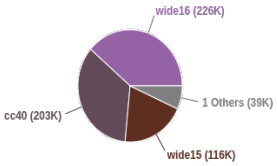
Domain	Docs	% of total
com	346K	59.18
net	36K	6.16
org	30K	5.12
com.ph	19K	3.30
ph	12K	2.09
pl	11K	1.80
nl	8.2K	1.41
info	7.2K	1.23
sg	6.5K	1.11
de	6.4K	1.10

Documents size (in segments)

<= 25 segments **7.82%** (46K documents)
> 25 segments **92.18%** (539K documents)

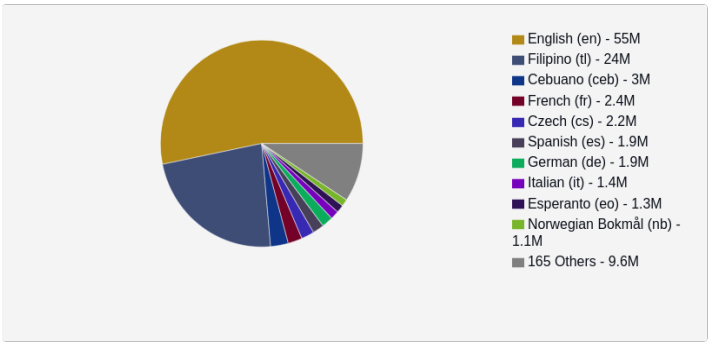


Documents by collection

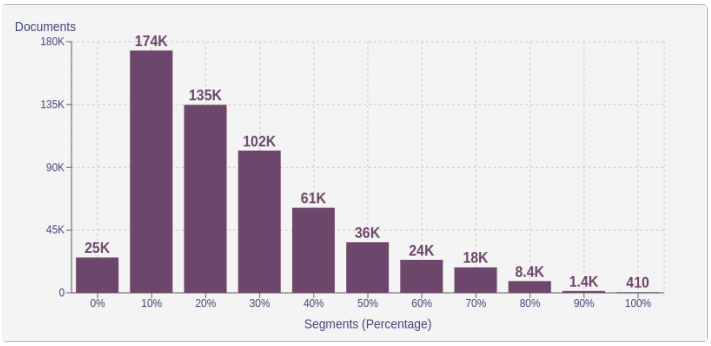


Language Distribution

Number of segments

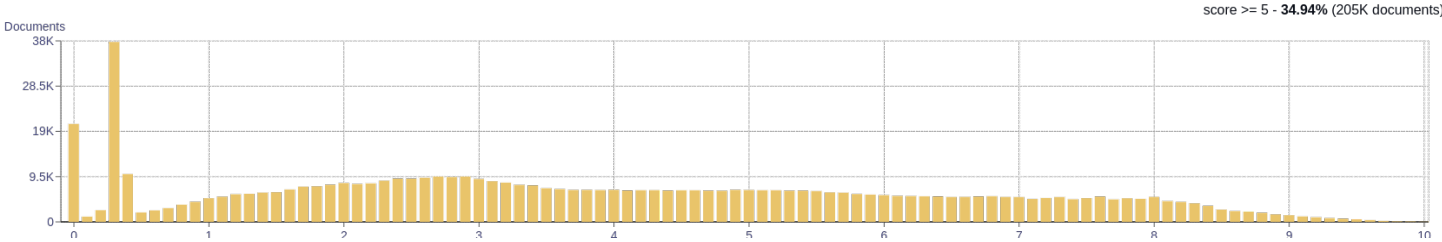


Percentage of segments in Filipino (tl) inside documents



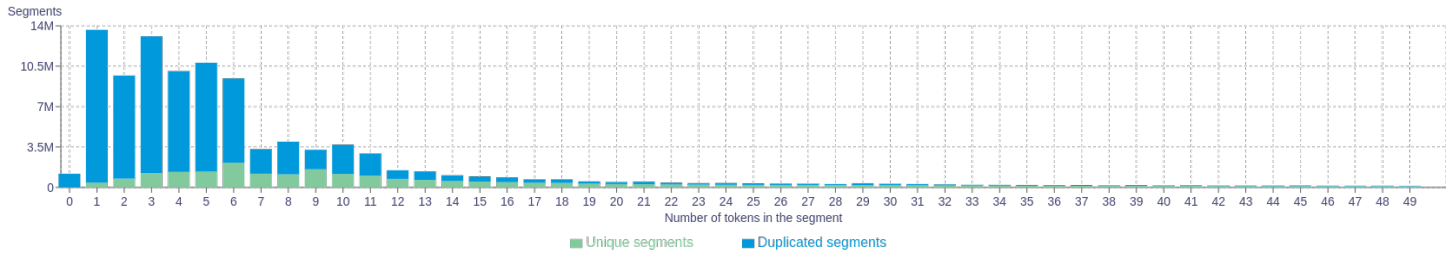
Distribution of documents by document score

score < 5 - **65.06%** (381K documents)
score >= 5 - **34.94%** (205K documents)

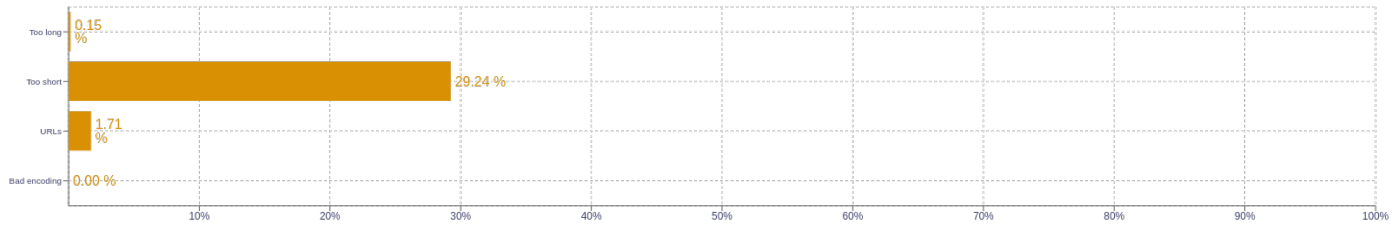


Segment length distribution by token

<= 49 tokens = **22M** segments | **78M** duplicates
> 50 tokens = **3.7M** segments | **1.1M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	the | 8231358 · to | 6859794 · and | 5257716 · of | 4599234 · code | 4322041
2	zip code | 1724206 · postal code | 1516374 · digit zip | 1028203 · -digit postal | 1027687 · last line | 981307
3	-digit postal code | 1027686 · address pangunahing numero | 980749 · zip code idagdag | 980627 · buliding firm pangalan | 980627 · address ikalawang number | 980627
4	id ng carrier ruta | 980626 · numero ng congressional district | 980616 · preferred last line key | 980612 · loob ng isang taon | 785608 · taon na ang nakalipas | 785435
5	accommodation photo ng accommodation photo | 139366 · share to twittershare to facebookshare | 126212 · twittershare to facebookshare to pinterest | 125672 · to twittershare to facebookshare to | 125672 · are you sure you want | 121837

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