HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
ceb_Latn.jsonl.tsv	9/6/2024	Cebuano (ceb)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
138,838	2,864,543	1,768,350 (61.73 %)	103M	493.16 MB	512,962,319

Top 10 domains

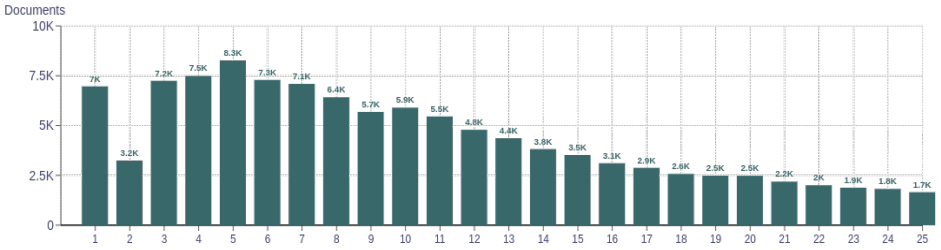
Domain	Docs	% of total
wikipedia.org	11K	7.91
jw.org	9.4K	6.75
blogspot.com	7K	5.01
sunstar.com.ph	5.3K	3.84
bible.is	4.8K	3.47
rmn.ph	3.8K	2.72
biblica.com	3.7K	2.67
philstar.com	3K	2.16
cleverplus.news	2.3K	1.66
eyewated.com	2.3K	1.65

Top 10 TLDs

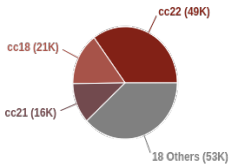
Domain	Docs	% of total
com	69K	49.86
org	29K	20.57
com.ph	7.5K	5.39
gov.ph	5.1K	3.70
is	4.9K	3.51
ph	4.6K	3.35
net	4K	2.89
news	2.4K	1.71
zone	1.6K	1.14
sg	1K	0.75

Documents size (in segments)

<= 25 segments **79.66%** (111K documents)
> 25 segments **20.34%** (28K documents)

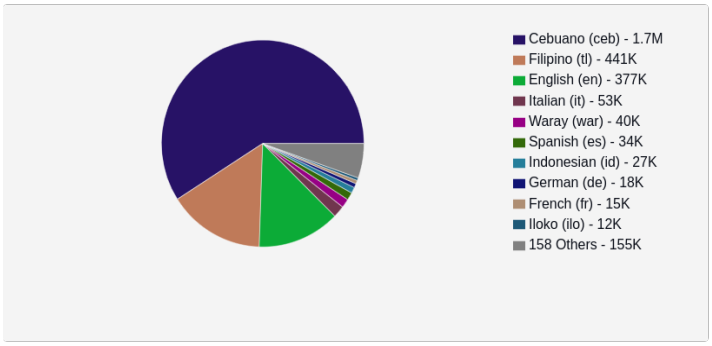


Documents by collection

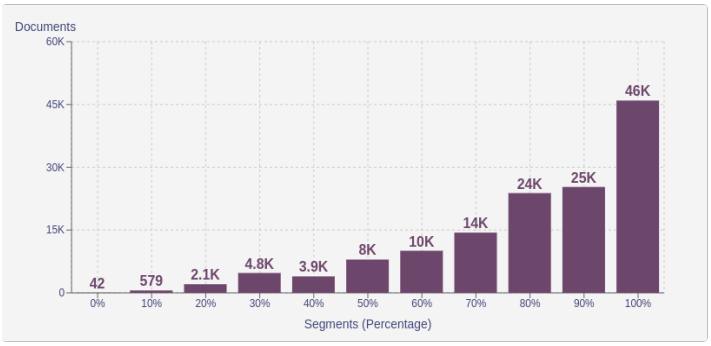


Language Distribution

Number of segments

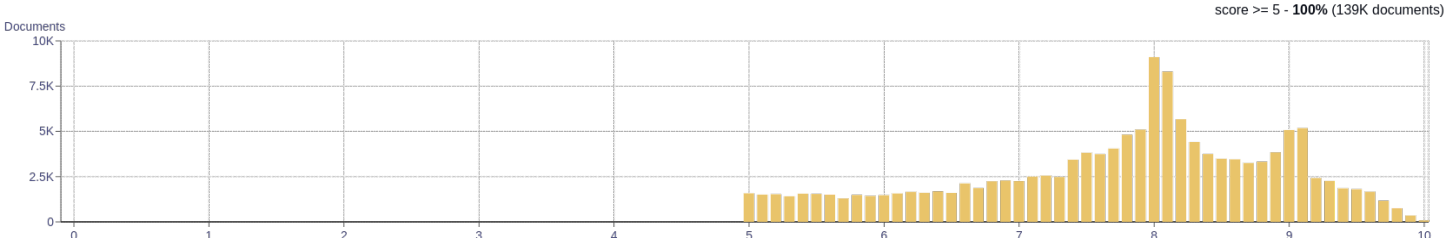


Percentage of segments in Cebuano (ceb) inside documents



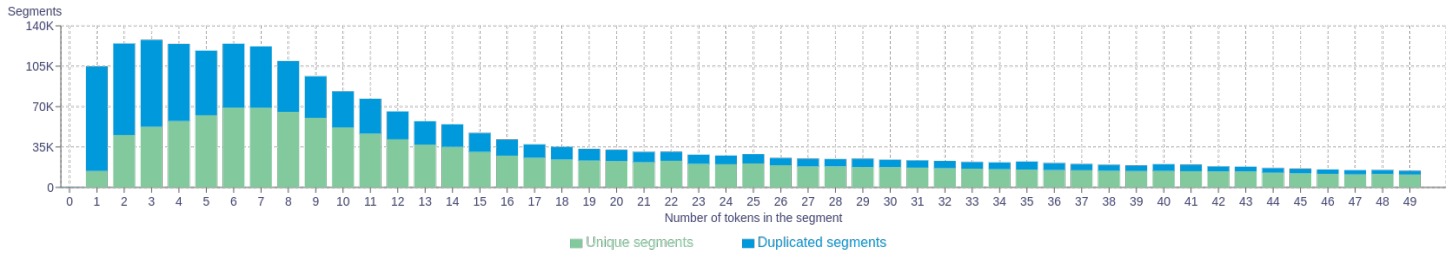
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (139K documents)

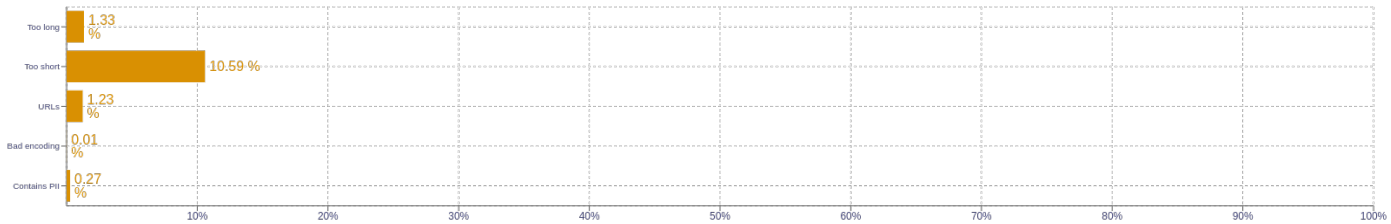


Segment length distribution by token

<= 49 tokens = **1.3M** segments | **923K** duplicates
> 50 tokens = **615K** segments | **173K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ka | 1288024 · usa | 770663 · ni | 500855 · na | 482791 · si | 475863
2	usa ka | 652793 · duha ka | 57652 · ni jehova | 42680 · ka tuig | 42307 · si jesus | 40573
3	adunay usa ka | 27769 · mao ang usa | 20398 · alang sa usa | 16321 · usa ka tawo | 15320 · anak nga lalake | 13596
4	mao ang usa ka | 16993 · alang sa usa ka | 15848 · anak nga lalake ni | 8158 · giklaro sa samang posisyon | 7026 · duha ka mga sumpay | 7026
5	sumpay sa giklaro sa samang | 7026 · ka mga sumpay sa giklaro | 7026 · sakop sa division nga ascomycota | 3838 · alang sa ubang mga dapit | 3677 · ubang mga dapit sa mao | 3675

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