

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

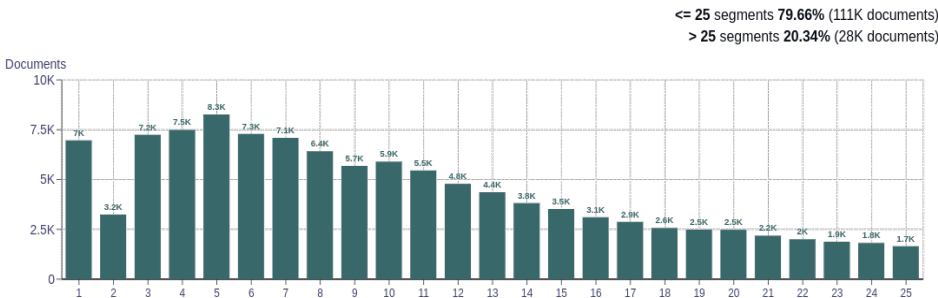
Top 10 domains

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

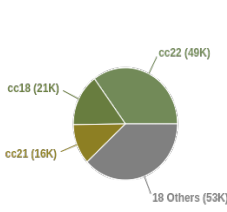
Top 10 TLDs

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)

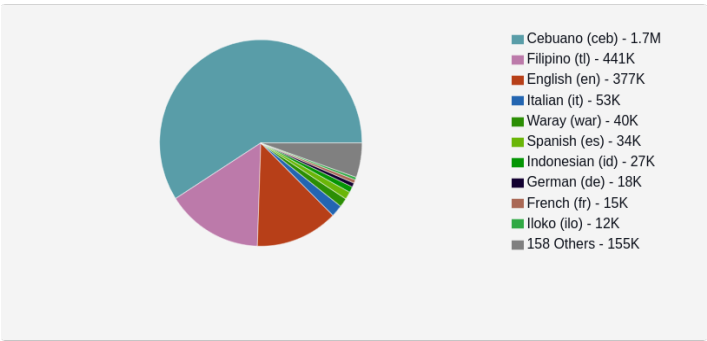


Documents by collection

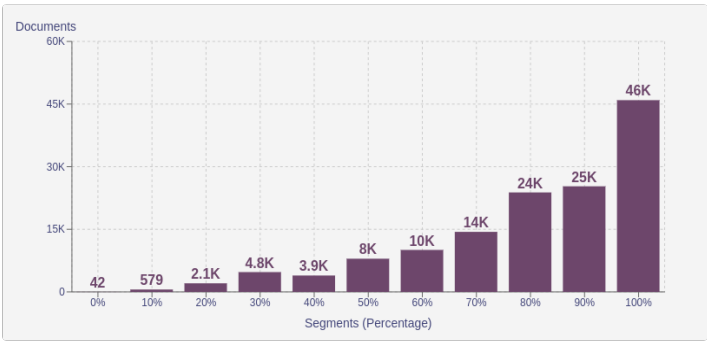


Language Distribution

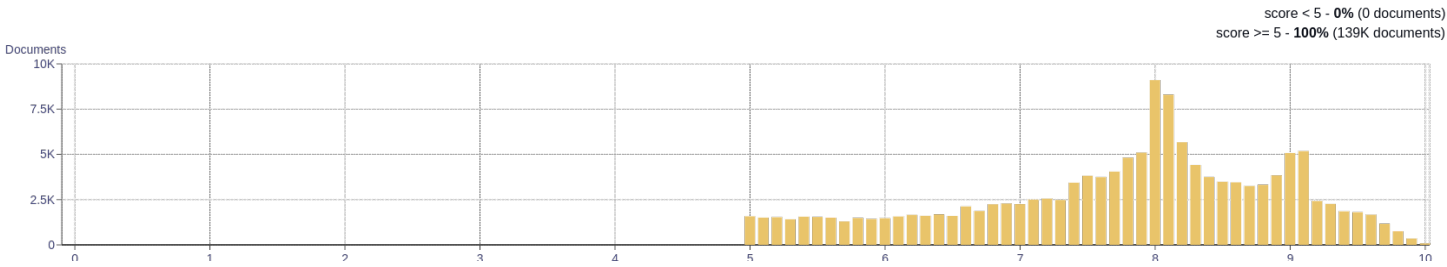
Number of segments



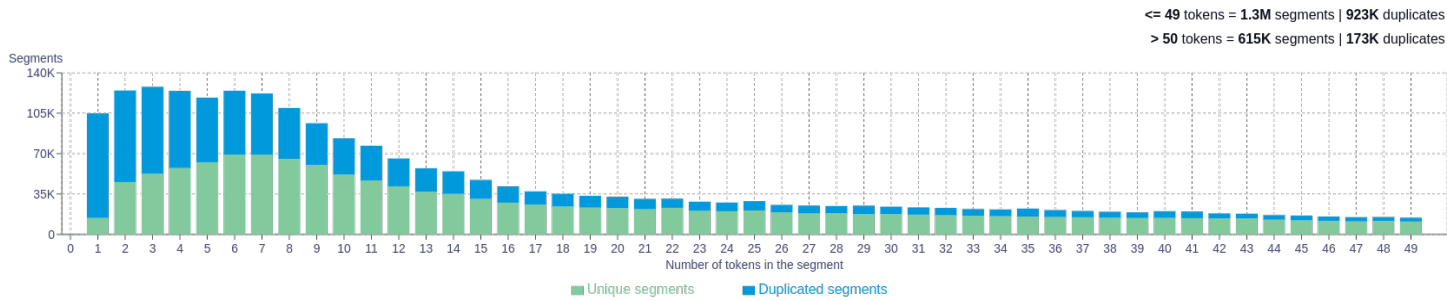
Percentage of segments in Cebuano (ceb) inside documents



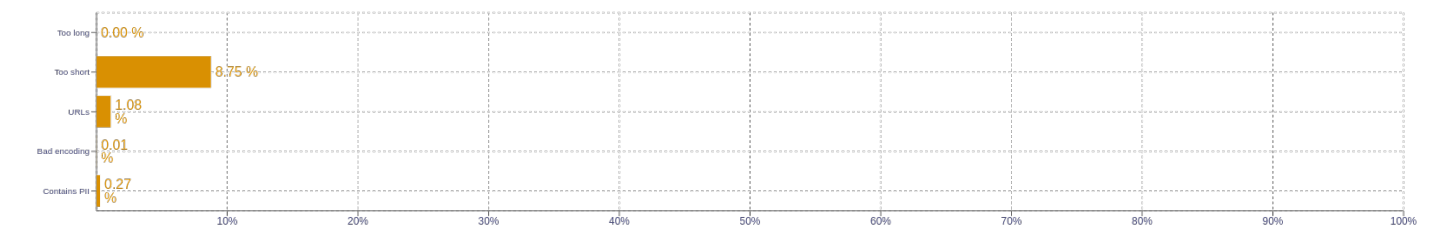
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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

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 of types (uniques)/number of tokens", after removing punctuation (<https://www.sitinfo.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>