

General overview

Corpus	Analytics date	Language
ttl_Latn.jsonl.tsv	9/6/2024	Filipino (tl)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,868,959	52,879,871	27,036,754 (51.13 %)	1.6B	7.57 GB	8,079,611,643

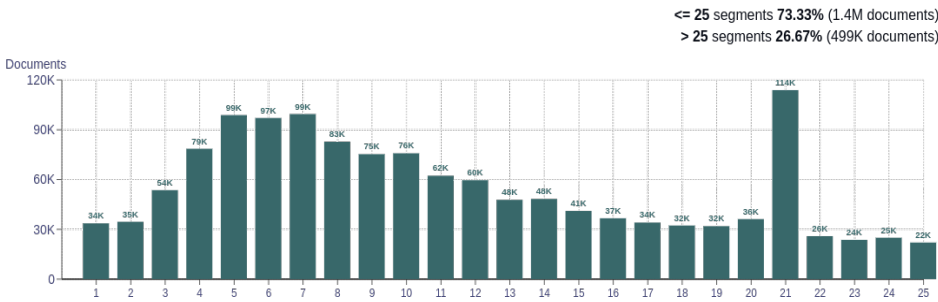
Top 10 domains

Domain	Docs	% of total
blogspot.com	119K	6.38
wikipedia.org	94K	5.04
remate.ph	77K	4.13
wordpress.com	51K	2.70
pinoyparazzi.com	35K	1.88
pep.ph	30K	1.61
dwiz882am.com	27K	1.46
abante.com.ph	25K	1.32
abs-cbn.com	23K	1.25
hatawtabloid.com	19K	1.02

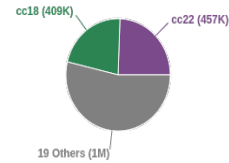
Top 10 TLDs

Domain	Docs	% of total
com	968K	51.81
org	197K	10.53
ph	163K	8.73
net	91K	4.89
ru	82K	4.37
com.ph	78K	4.15
pl	21K	1.13
gov.ph	21K	1.10
info	19K	1.04
tk	18K	0.97

Documents size (in segments)

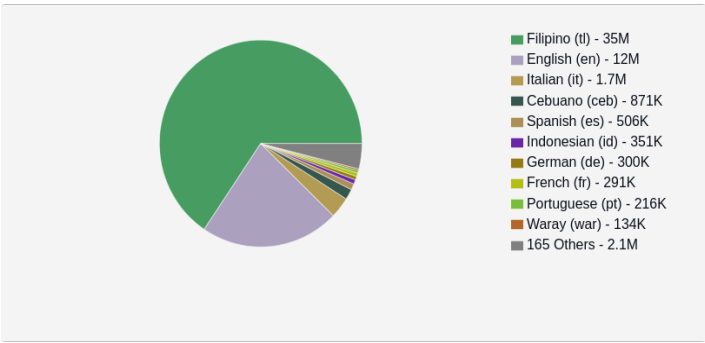


Documents by collection

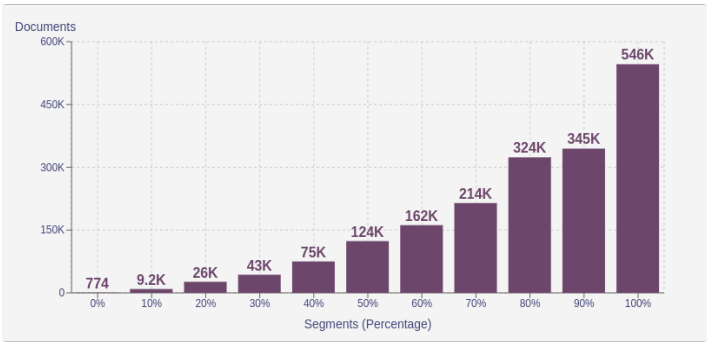


Language Distribution

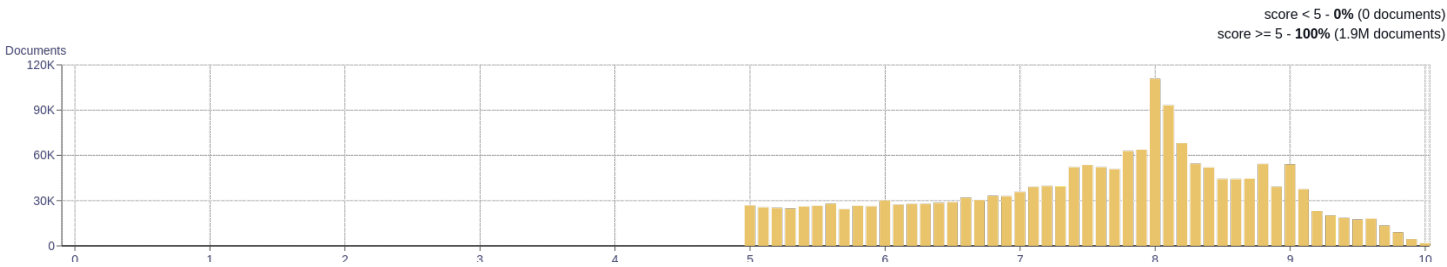
Number of segments



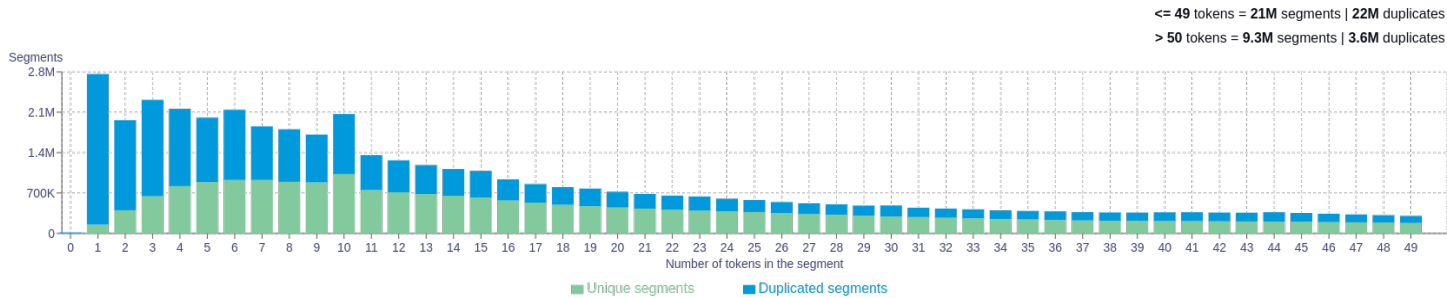
Percentage of segments in Filipino (tl) inside documents



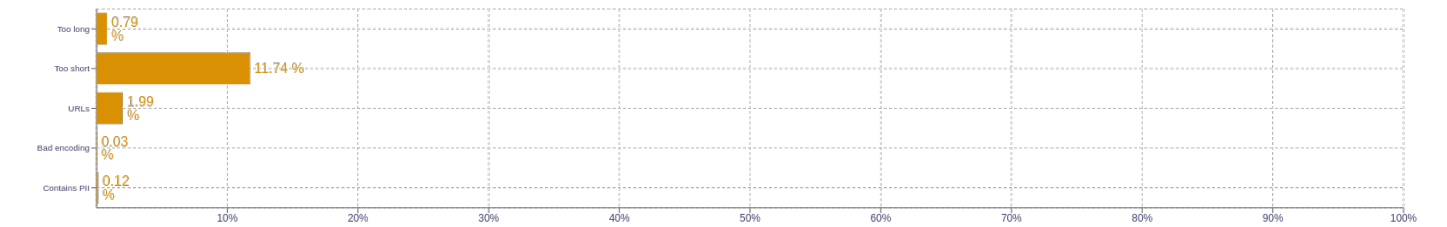
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>si 5734355</div> <div>mo 5100988</div> <div>the 4926297</div> <div>lang 4737829</div> <div>naman 4033321</div>
2	<div>of the 635409</div> <div>kuko halamang-singaw 386309</div> <div>in the 383085</div> <div>t ibang 293785</div> <div>'t ibang 251634</div>
3	<div>pagbaba ng timbang 221070</div> <div>pati na rin 174092</div> <div>mawalan ng timbang 150615</div> <div>you must be 139435</div> <div>logged in to 138324</div>
4	<div>you must be logged 138211</div> <div>must be logged in 138210</div> <div>be logged in to 138209</div> <div>to post a comment 137377</div> <div>in to post a 137336</div>
5	<div>you must be logged in 138209</div> <div>must be logged in to 138200</div> <div>in to post a comment 137326</div> <div>logged in to post a 137311</div> <div>be logged in to post 137308</div>

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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.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>, , , 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>