

General overview

Corpus	Analytics date	Language
yue_Hant.jsonl.tsv	9/6/2024	Cantonese (yue)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
61,286	1,235,188	799,854 (64.76 %)	46M	179.89 MB	73,128,238

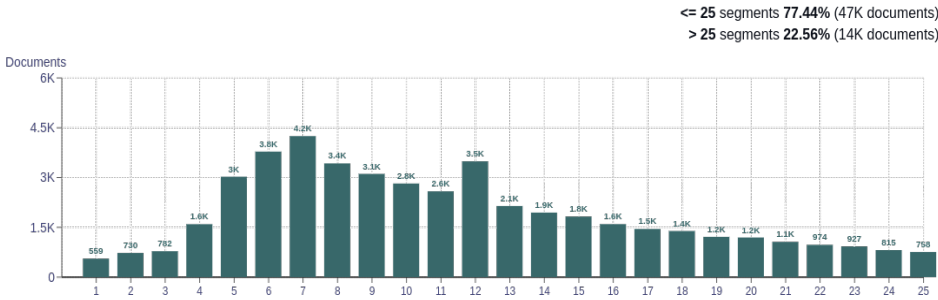
Top 10 domains

Domain	Docs	% of total
blogspot.hk	7.8K	12.78
wikipedia.org	4.2K	6.83
vjmedia.com.hk	4K	6.46
blogspot.com	2.7K	4.43
kennethkwok.me	1.6K	2.63
blogspot.tw	1.3K	2.18
hotels.com	1.2K	2.01
blogspot.sg	1.2K	1.99
hkgolden.com	1.2K	1.98
collection.news	1.2K	1.94

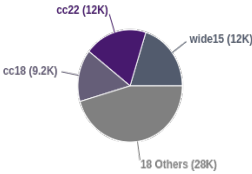
Top 10 TLDs

Domain	Docs	% of total
com	26K	42.60
hk	11K	18.35
com.hk	7.5K	12.21
org	4.8K	7.88
me	1.8K	2.98
net	1.6K	2.55
tw	1.4K	2.26
news	1.2K	2.00
sg	1.2K	1.99
name	824	1.34

Documents size (in segments)

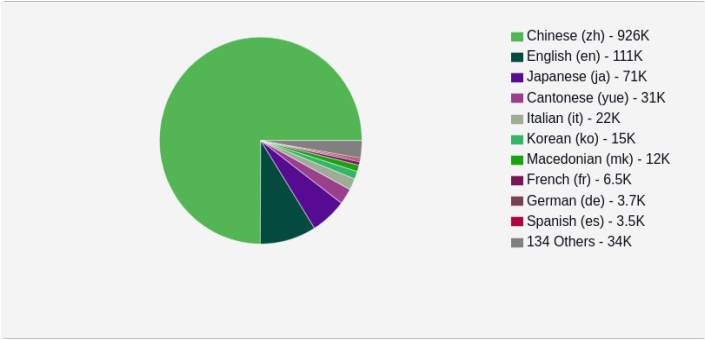


Documents by collection

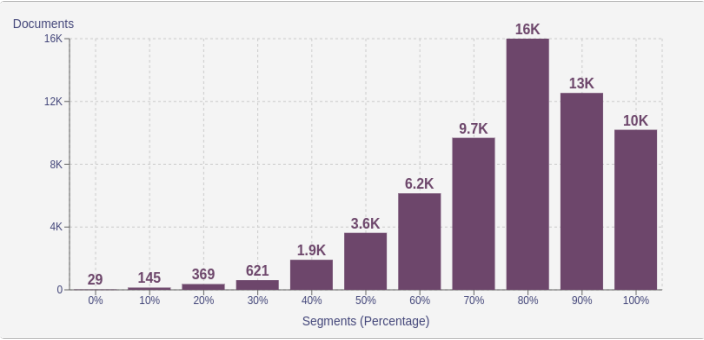


Language Distribution

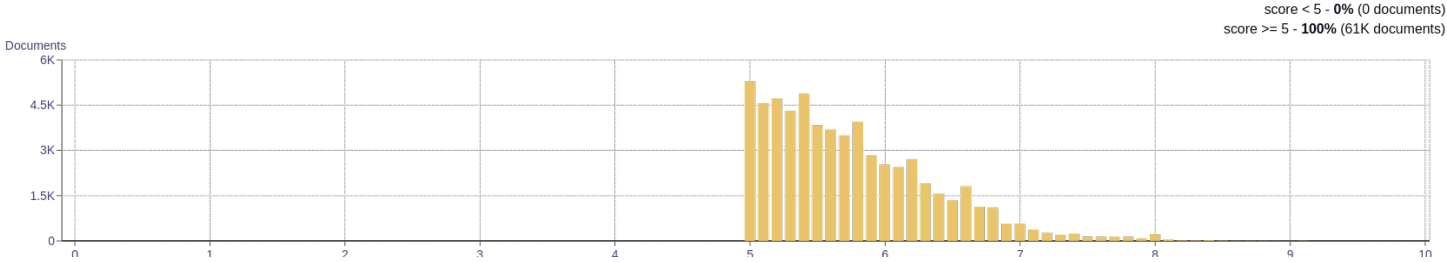
Number of segments



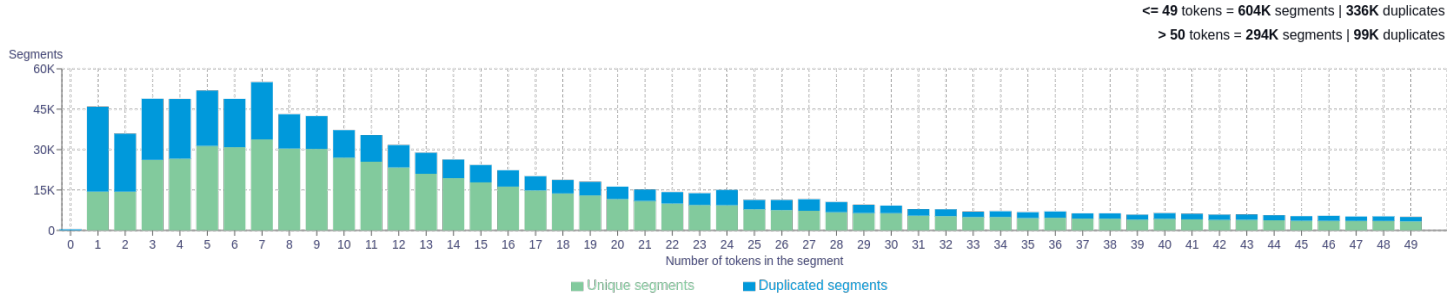
Percentage of segments in Cantonese (yue) inside documents



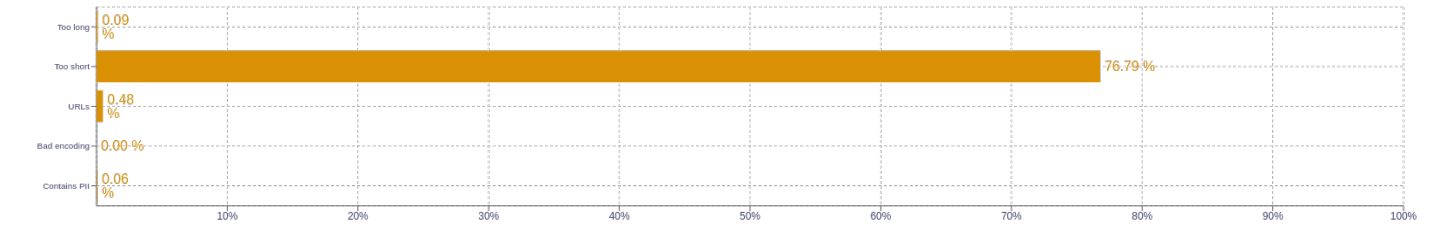
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>既 198369</div> <div>人 160146</div> <div>啦 147086</div> <div>地 146371</div> <div>左 142078</div>
2	<div>讀 讀 52763</div> <div>好好 好好 32528</div> <div>娛 乐 24828</div> <div>玩 家 汇 24682</div> <div>汇 娛 24682</div>
3	<div>讀 讀 讀 40845</div> <div>玩 家 汇 娛 24682</div> <div>汇 娛 乐 24682</div> <div>好好 好好 好好 21259</div> <div>廢 物 廢 物 廢 物 13824</div>
4	<div>讀 讀 讀 讀 29460</div> <div>玩 家 汇 娛 乐 24682</div> <div>好好 好好 好好 好好 15062</div> <div>廢 物 廢 物 廢 物 廢 物 13821</div> <div>慢慢 慢慢 慢慢 慢慢 13225</div>
5	<div>讀 讀 讀 讀 讀 19456</div> <div>廢 物 廢 物 廢 物 廢 物 廢 物 13818</div> <div>慢慢 慢慢 慢慢 慢慢 慢慢 13172</div> <div>版 app- 玩 家 汇 娛 12332</div> <div>app- 玩 家 汇 娛 乐 12332</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.stinfo.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>