HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

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

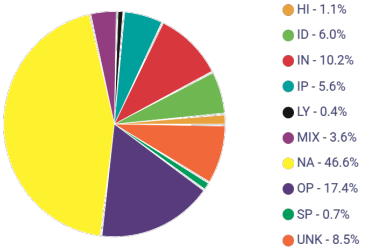
Top 10 domains

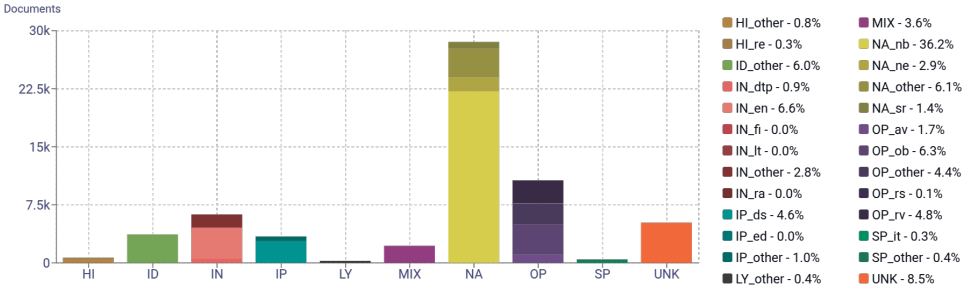
Domain	Docs	% of total
blogspot.hk	7.8K	12.78%
wikipedia.org	4.2K	6.83%
vjmedia.com.hk	4K	6.46%
blogspot.com	2.7K	4.43%
kennethkwok.me	1.6K	2.63%
blogspot.tw	1.3K	2.18%
hotels.com	1.2K	2.01%
blogspot.sg	1.2K	1.99%
hkgolden.com	1.2K	1.98%
collection.news	1.2K	1.94%

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%

Register labels

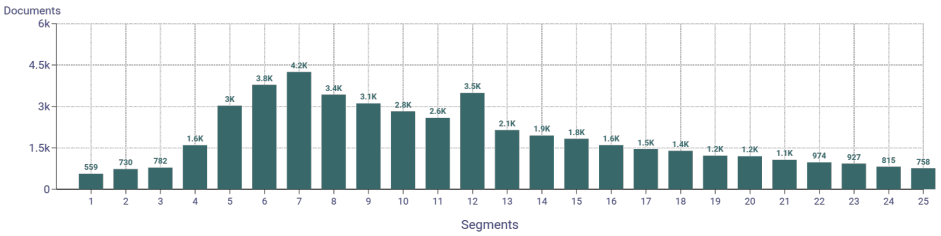




MT:0.5% | 332 Documents

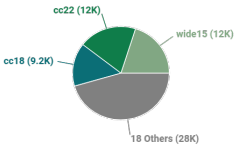
Documents size (in segments)

<= 25 segments **77.44%** (47K documents)
> 25 segments **22.56%** (14K documents)



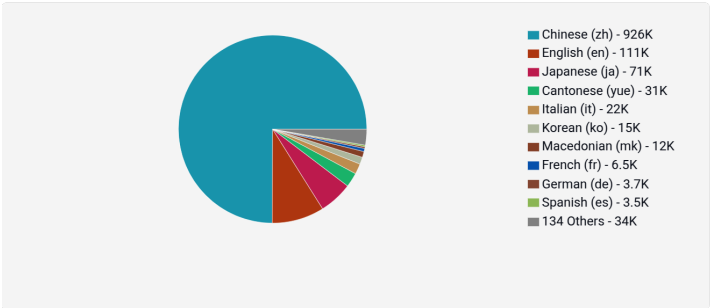
Documents by collection

CC = 49.43%
IA = 50.57%

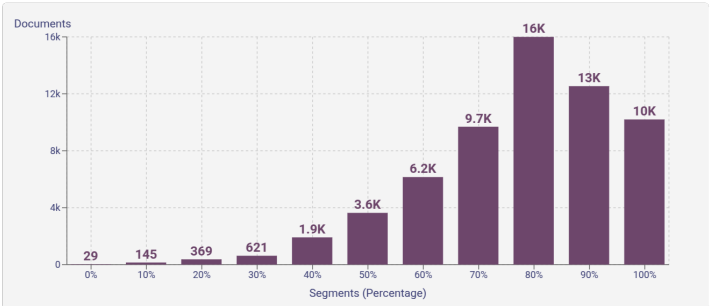


Language Distribution

Number of segments in the Cantonese (yue) corpus

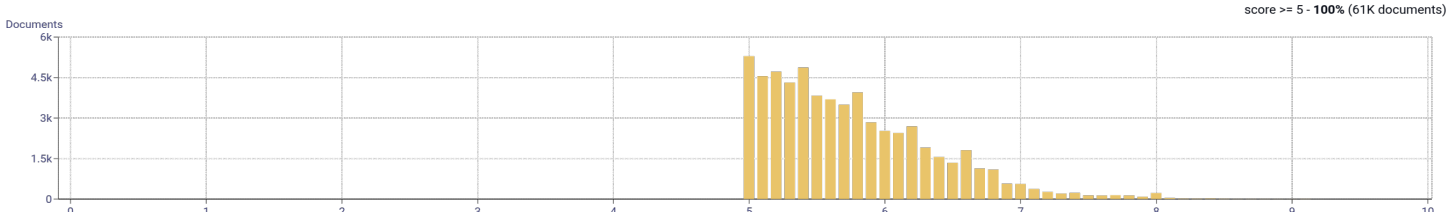


Percentage of segments in Cantonese (yue) inside documents



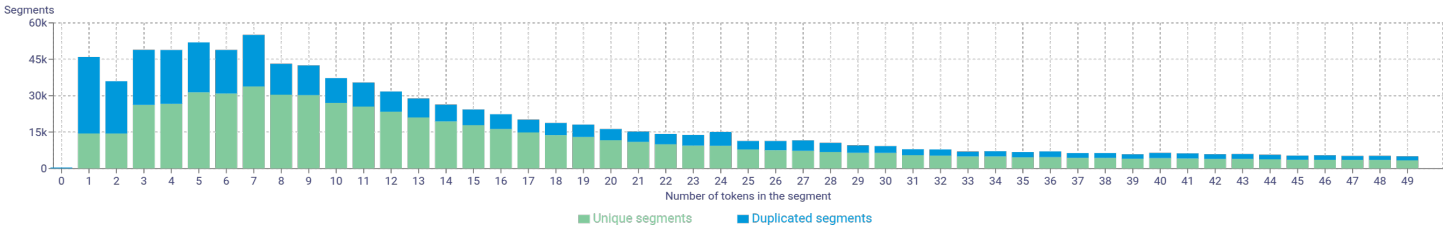
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (61K documents)

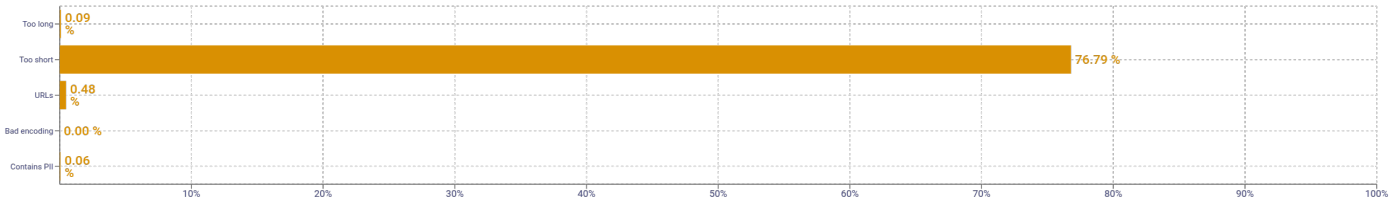


Segment length distribution by token

≤ 49 tokens = **604K** segments | **336K** duplicates
> 50 tokens = **294K** segments | **99K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	既 | 198369 · 人 | 160146 · 啦 | 147086 · 地 | 146371 · 左 | 142078
2	讀 讀 | 52763 · 好好 好好 | 32528 · 娛 乐 | 24828 · 玩家 汇 | 24682 · 汇 娛 | 24682
3	讀 讀 讀 | 40845 · 玩家 汇 娛 | 24682 · 汇 娛 乐 | 24682 · 好好 好好 好好 | 21259 · 廢物 廢物 廢物 | 13824
4	讀 讀 讀 讀 | 29460 · 玩家 汇 娛 乐 | 24682 · 好好 好好 好好 好好 | 15062 · 廢物 廢物 廢物 廢物 | 13821 · 慢慢 慢慢 慢慢 慢慢 | 13225
5	讀 讀 讀 讀 讀 | 19456 · 廢物 廢物 廢物 廢物 廢物 | 13818 · 慢慢 慢慢 慢慢 慢慢 慢慢 | 13172 · 版 app- 玩家 汇 娛 | 12332 · app- 玩家 汇 娛 乐 | 12332

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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av