



# Term Project1

Team 3

**Hosted By:**

A1105505 林彧頤

A1105507 蘇柏諺

A1105521 黎子歲

A1105523 巫柔筠

A1105524 吳雨宣

**IR & GenAI**

# Table of Contents

- 01** Multi-source Crawling
- 02** Real-time Data Collection
- 03** Topic-specific Crawling
- 04** Sentiment Analysis
- 05** Dynamic and Adaptive Crawling
- 06** Data Storage
- 07** Visualization

**1.**

# **Multi-source Crawling**



	中時	自由時報	tvbs	API								
時間	最多爬到2年前 (2022/10/24)	最多只能爬500頁 (有些只能爬不到2年， 有些最久可爬到2004年)	最多只能爬295頁 (2003/6/3)	最多爬到1個月前								
資料 筆數	H	St	Sp	H	St	Sp	H	St	Sp	H	St	Sp
	5068	11903	15895	15268	19188	13894	15952	16013	30422	237	373	351
	32866			48350			62387			961		
144564												
新聞 來源數	1			1			1			24		

# 中時

- **模擬 User-Agent:** 隨機選擇 User-Agent，模擬不同裝置訪問。
- **隨機延遲:** 每次請求隨機延遲1-3秒。
- **隱藏WebDriver:** 將 `navigator.webdriver` 設為 `undefined`。
- **多組選擇器:** 使用 `safe\_find\_element` 指定多組 CSS 選擇器。
- **重試機制:** 若找不到元素，最多重試3次。
- **頁面重新打開:** 抓取一頁後關閉瀏覽器並重新打開。

# 自由時報

- **延遲請求:** 每爬取一頁延遲1秒，避免頻繁請求。
- **例外處理與重載:**  
使用 WebDriverWait 和 TimeoutException，超時則重新載入頁面。
- **過濾重複連結:**  
使用 `seen\_hrefs` 集合儲存重複的新聞連結，避免重複爬取。

- **延遲請求:** 每頁爬取後延遲2秒，避免異常行為。
- **多重選取器容錯:** 使用多選取器定位，增強兼容性。
- **異常處理:**  
用 try-except 忽略單篇解析錯誤，確保爬蟲不會因為小問題而停止。

**2.**

# **Real-time Data Collection**



# 實時數據收集

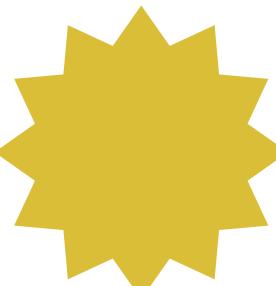
- 持續爬取新文章:

爬蟲會從新聞網第一頁開始抓取最新發布的文章，確保資料即時更新。

- 資料儲存設計(Supabase):

支援即時數據同步，方便組員同時操作數據。

免費版單一資料庫容量上限為500MB，因為我們資料龐大，創建了兩個資料庫。



# 資料結構與索引優化

- 資料庫結構:

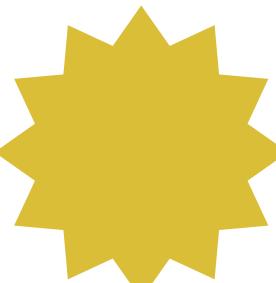
每篇新聞保存 title, source, date, URL, 及 content

- 全文檢索優化:

content 欄位儲存大量文本資料，進行全文檢索時容易降低查詢效率。

```
CREATE INDEX sport_news_content_gin_idx  
ON sport_news USING GIN (to_tsvector('simple', content));
```

使用 GIN 索引來加速全文檢索操作(stock\_news, health\_news, sport\_news )



# 3. Topic-specific Crawling



# Crawling Strategy

中時 新聞網 真道理性 真愛台灣

即時 政治 言論 生活 娛樂 財經 股市 國際 兩岸 軍事

76歲也能

首頁 / 搜尋結果

搜尋結果

以下是含有 **股市** 的搜尋結果，共 **19,132** 筆

單一以主題為關鍵字搜尋  
(資料量少)

真道理性 真愛台灣

中時 新聞網

即時 政治 言論 生活 娛樂 財經

首頁 / 搜尋結果

**搜尋結果**

以下是含有 **台積電** 的搜尋結果，共 **21,775** 筆

設定出單一類別就會比大方向的資料來的多

主題下多類別為關鍵字搜尋  
(資料量多)

# Crawling Strategy-cont.



記憶已更新

我會建議你以以下10個子題進行健康主題的情緒分析：

1. 確診 - 涉及疾病的確診結果，通常帶有情緒波動。
2. 減重 - 關於體重管理，與健康生活方式相關，常引發正面或焦慮情緒。
3. 免疫力 - 討論提升或保持免疫系統健康的話題，特別是在流行病時期。
4. 營養 - 營養攝取和膳食平衡，對健康情緒的影響頗大。
5. 疾病 - 涵蓋不同疾病的討論，情緒上可能複雜多樣。
6. 病毒 - 尤其是疫情相關的病毒話題，通常帶有擔憂或焦慮的情緒。
7. 慢性病 - 長期病症的管理與影響，涉及患者和照護者的情感變化。
8. 健康飲食 - 營養均衡和飲食選擇的情緒分析，與生活品質緊密相連。
9. 癌症 - 這類話題對病患和家屬的情緒影響深刻，常帶有嚴肅的情感基調。
10. 醫療 - 涉及醫療品質、診斷準確性等，對患者的信任和滿意度有直接影響。

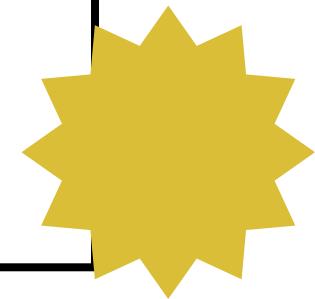
這些子題既涵蓋了具體健康問題，也涉及預防和生活方式等不同方面，對情緒分析的完整性和深度很有幫助。



ChatGPT來回答  
最好的十個子題做參考  
並人工做增修刪

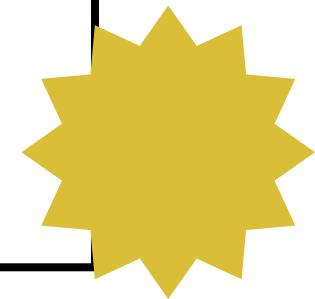
# 主題與子題含意-Stock

統一	追蹤統一企業於市場的營運狀況與投資機會。	聯發科	探究聯發科在半導體市場的產品競爭力。
鴻海	了解鴻海供應鏈及全球市場的科技影響力。	長榮	觀察長榮於運輸和物流業的業績走勢。
台積電	探索台積電於半導體領域的競爭優勢。	富邦金	了解富邦金於金融保險市場的佈局策略。
廣達	分析廣達電子製造業務的成長趨勢與合作夥伴。	長榮航	分析長榮航空的航線拓展及旅遊趨勢。
中華電	追蹤中華電在電信業的發展和技術創新。	元大金	研究元大金於資產管理及金融市場的角色。



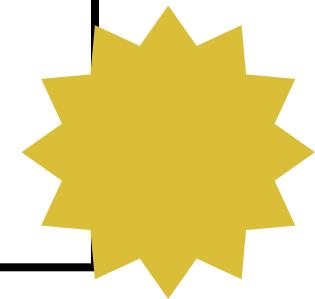
# 主題與子題含意-Health

確診	追蹤疫情確診數據以評估健康風險。	病毒	研究病毒傳播途徑與防護措施。
減重	探索減重方法及其對健康的影響。	慢性病	了解慢性病的成因及健康管理策略。
免疫力	了解增強免疫力的方法和相關保健品。	健康飲食	介紹健康飲食的要素與對身體的益處。
營養	分析營養對健康的益處與飲食建議。	癌症	探索癌症的最新治療和預防進展。
疾病	探討常見疾病的預防和治療方法。	醫療	追蹤醫療科技發展及對健康的影響。



# 主題與子題含意-Sports

賽事	追蹤各類運動賽事的賽程及結果。	奧運	研究奧運賽事及選手表現的亮點。
球員	研究球員的表現和職業生涯發展。	NBA	探索NBA的明星球員與聯賽動態。
聯賽	分析各大聯賽的競爭趨勢及影響力。	MLB	了解MLB的賽季變化及選手動向。
世界盃	探討世界盃比賽及球隊的參賽表現。	棒球	研究棒球運動的發展和國際賽事。
籃球	追蹤籃球比賽及其全球影響力。	中職	追蹤中華職棒的賽事及球員動態。



# 處理不相關的資料

- 爬蟲中使用關鍵字
- Gemini prompt 分析過濾(如下)
- Gemini的API不可用->keyword檢索

```
prompt = f"""
```

以下是今天的新聞資訊。請根據這些資訊判斷其是否是股市新聞，如果不是請回答無關：

標題：

{title}

文章內容：

{news}

....

# 4. Sentiment Analysis



# 資料蒐集與預處理

## 資料來源

- 從 Supabase 提取新聞數據，使用批量查詢以提高效率。

## 資料清理

固定輸入長度：模型對輸入文本的長度有上限 => 最大字元數為 512。超過長度的文本將無法被模型有效處理，可能導致信息丟失或模型無法返回有效的預測結果。

文本內容截斷：確保輸入的新聞內容可以完整地傳遞給模型

=> 選擇前 512 字元通常可以保留關鍵的上下文信息

=> 提高計算效率，限制輸入長度可以加快模型的計算速度

=> 確保一致性：統一的輸入長度有助於減少模型在訓練和推理階段的不一致性

# BERT 情緒分析模型

"nlptown/bert-base-multilingual-uncased-sentiment"

由 NLP Town 提供的多語言情緒分析模型，基於 BERT 架構進行訓練。  
多語言支持，能夠處理不同語言的文本，包括中文。

## 使用場景

社交媒体情緒分析、產品評論分析、新聞情緒評估

## 技術特點

通過標記不同情緒的多語言文本數據進行預訓練的，具備一定的通用性，但在特定領域中可能需要進行微調。

# 模型使用方法

模型載入：通過 Hugging Face Transformers 庫加載

# 配置情緒分析模型 (使用 nlptown/bert-base-multilingual-uncased-sentiment 模型)

```
model_name = "nlptown/bert-base-multilingual-uncased-sentiment"  
model = BertForSequenceClassification.from_pretrained(model_name)  
tokenizer = BertTokenizer.from_pretrained(model_name)
```

# 創建情緒分析 pipeline

```
sentiment_analyzer = pipeline("sentiment-analysis", model=model,  
tokenizer=tokenizer)
```

# 資料分析與計算

## 分析邏輯

### (1) 星級計數：

根據每則新聞的評分星級進行分類，生成星級分佈。

模型輸出從 1 到 5 的分數來表示情緒強度。



非常負面

中立

非常正面

# 資料分析與計算

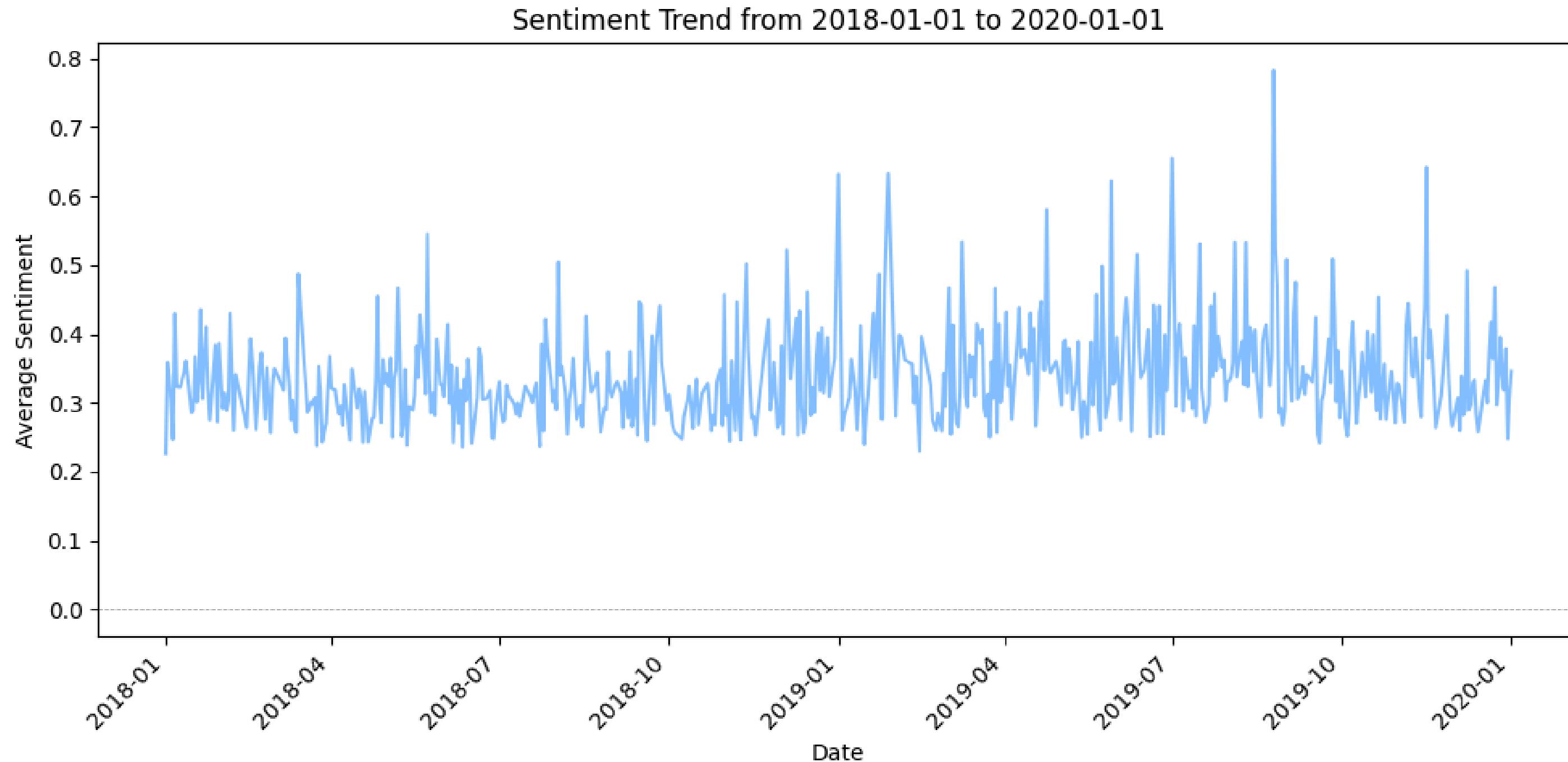
## (2) 情緒計數：

逐步累加新聞中出現的正面、中立及負面情緒

# 根據星級設定情緒類型

```
if star in [1, 2]:  
    emotion = -1 # negative  
elif star == 3:  
    emotion = 0 # neutral  
else:  
    emotion = 1 # positive
```

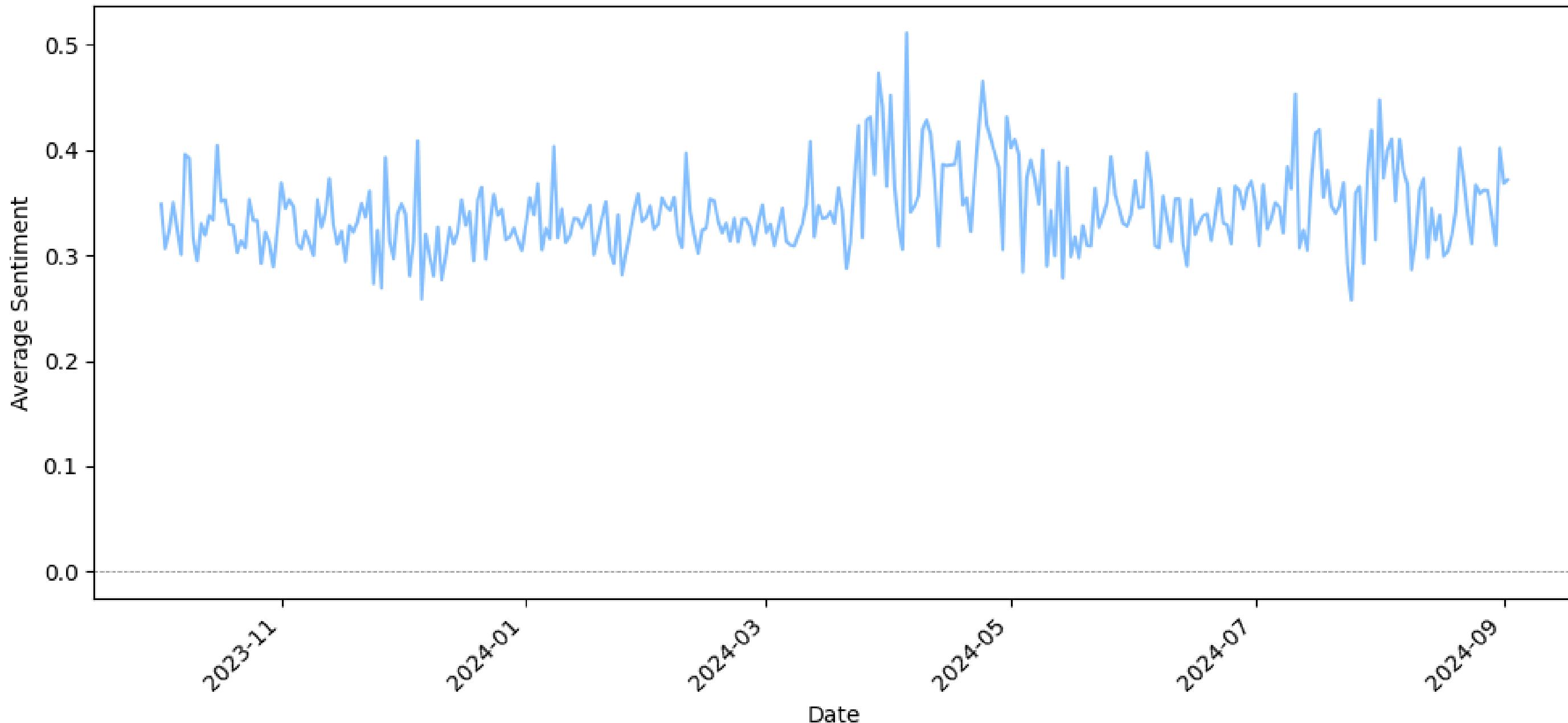
#1



這是health 2018~2020，由此顯示人民對健康的相關議題愈來愈有信心，因為技術的上升，其中2019年是分界點，因為在疫情下的台灣，展現出極高的防禦心理，造成新聞多為正面報導

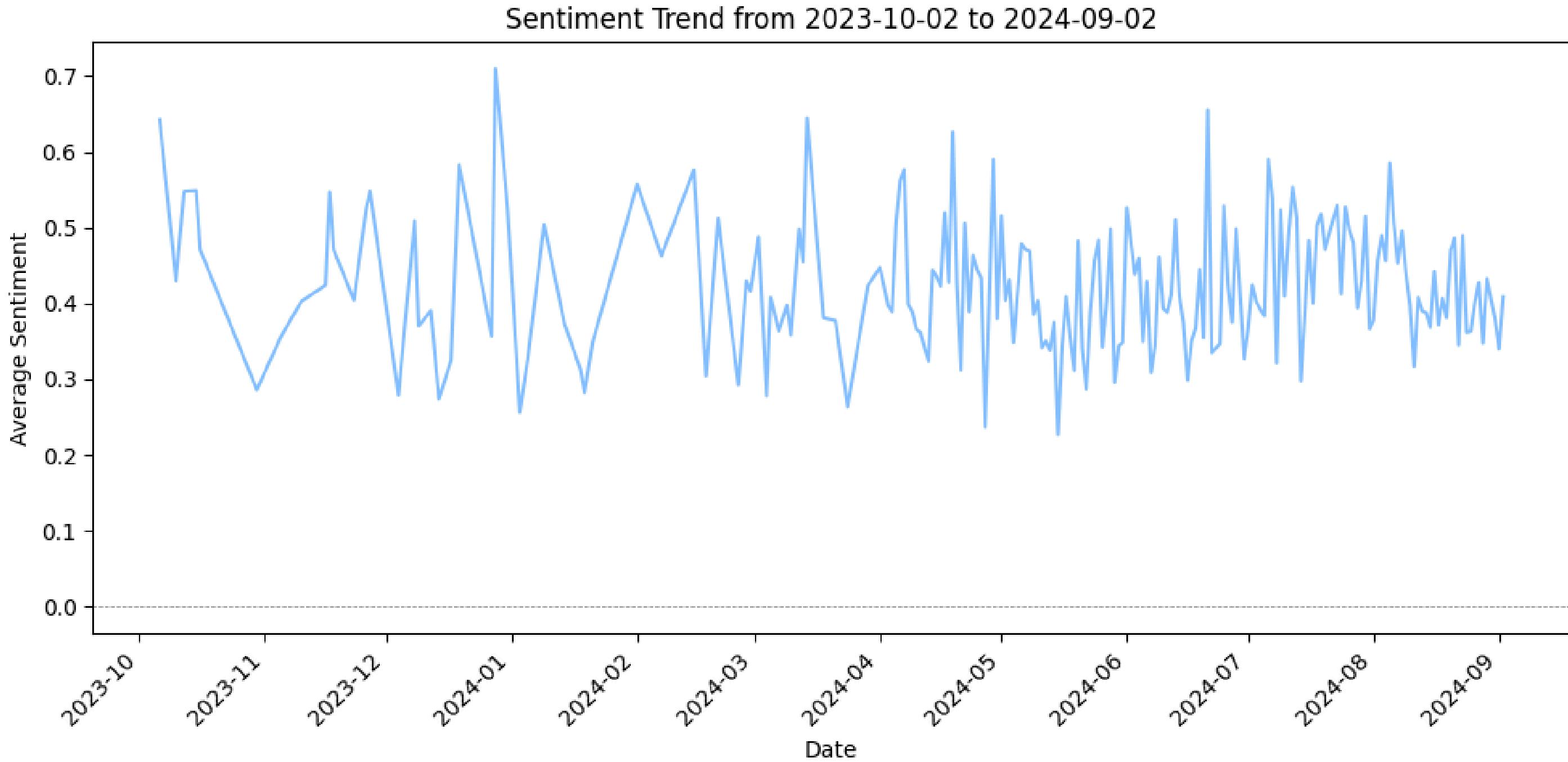
#2

Sentiment Trend from 2023-10-02 to 2024-09-02



由此圖可以顯示，由於近2024年奧運，因此有關於運動的相關議題，  
皆於2024開始情緒上漲，並於2024 8月屬於高峰值群

#3



股市方面，可以發現是屬於動盪較大的情緒指標，這會與當時刻的股市大盤走向有關，而最近因為戰爭、天災等等...因素，因此情緒指標走跌

**5.**

# **Dynamic and Adaptive Crawling**



# Dynamic Crawling

- **Selenium**：自動化操作瀏覽器。
- **BeautifulSoup**：解析靜態 HTML。
- **API**：直接使用網站 API 更穩定高效，省去網頁渲染和加載。
- **User Agent 和 Headers**：模擬真實瀏覽器，避免封鎖。根據網站需求設置合適的 User Agent 和 Headers。
- **模擬人類行為**：隨機停頓與動態操作，防止被網站識別為爬蟲。
- **生成式 AI**：輔助過濾新聞並提供爬蟲調整建議，應對網站架構變動。

# Adaptive Crawling

- **CSS Selector**: 透過 CSS 選擇器定位元素，適應結構變動，降低中斷風險。
- 正則表達式：提取關鍵資料，避免依賴 HTML 結構，提高代碼靈活性。
- 類似元素辨識：結構變動時，根據標籤或屬性模式找到相似元素。
- **User Agent**：根據網站需求更改 User Agent，應對反爬機制。
- 生成式 AI：協助撰寫與優化爬蟲代碼，適應網站結構變化。
- 錯誤通知與重試：異常時自動通知與重試，減少數據缺失。

# 實際證明

放置於server不間斷跑了2星期，皆無不成功爬蟲！

可以動態適應改變的結構，爬蟲機器被視為人為操作~

# 6. Data Storage



# stock

stockID	stock_name
股票代號	股票名稱

# stock\_news

stockID	title	date	content	source	id	url
股票代號	標題	日期	內容	來源	index	連結

# stock\_news\_API

stockID	title	date	content	source	id	url
股票代號	標題	日期	內容	來源	index	連結

# stock\_news\_sentiment

id	news_id	sentiment	emotion	star
index	新聞編號	情緒分數	情緒類型	星等

# health

healthID	health_name
keyword代號	keyword名稱

# health\_news

healthID	title	date	content	source	id	url
keyword代號	標題	日期	內容	來源	index	連結

# health\_news\_API

healthID	title	date	content	source	id	url
keyword代號	標題	日期	內容	來源	index	連結

# health\_news\_sentiment

id	news_id	sentiment	emotion	star
index	新聞編號	情緒分數	情緒類型	星等

# **sport**

<b>sportID</b>	<b>sport_name</b>
keyword代號	keyword名稱

# **sport\_news**

<b>sportID</b>	<b>title</b>	<b>date</b>	<b>content</b>	<b>source</b>	<b>id</b>	<b>url</b>
keyword代號	標題	日期	內容	來源	index	連結

# **sport\_news\_API**

<b>sportID</b>	<b>title</b>	<b>date</b>	<b>content</b>	<b>source</b>	<b>id</b>	<b>url</b>
keyword代號	標題	日期	內容	來源	index	連結

# **sport\_news\_sentiment**

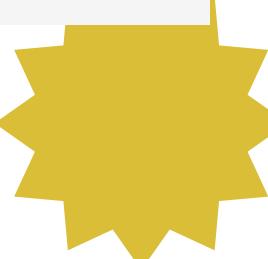
<b>id</b>	<b>news_id</b>	<b>sentiment</b>	<b>emotion</b>	<b>star</b>
index	新聞編號	情緒分數	情緒類型	星等

# Querying articles

(1) 資料以date排序

```
SELECT *
FROM health_news
ORDER BY date ASC;
```

	id int	title varchar	date date	content varchar	source varchar	url varchar	id int
	15458	六旬老翁瘦成功 不褪高壓電流	2003-06-05	台中縣有一位60歲的民衆，十多年前單了氣功後，就發現自己	tvbs	<a href="https://news.tvbs.com.tw/life/412668">https://news.tvbs.com.tw/life/412668</a>	15458
	16306	羅原病尋S疾難 痘醫局：容易感染SARS	2003-06-07	陽明醫院爆發SARS院內群聚感染，不過衛生署表示並不影響	tvbs	<a href="https://news.tvbs.com.tw/life/412412">https://news.tvbs.com.tw/life/412412</a>	16306
	16306	67歲老翁瘦成功 情況各色指管	2003-06-07	11名陽明醫學院SARS爆發患者，恢復要達到松山醫學院，才	tvbs	<a href="https://news.tvbs.com.tw/life/412382">https://news.tvbs.com.tw/life/412382</a>	16306
	16304	陽明11名SARS病患 2患者病危	2003-06-08	陽明醫學院爆發的11名SARS病患，又有2名病患列為可能病危	tvbs	<a href="https://news.tvbs.com.tw/life/412330">https://news.tvbs.com.tw/life/412330</a>	16304
	15457	上海體育局：瘦弱老翁瘦成功	2003-06-10	上海市的SARS疫情雖然比較和緩，不過上海市政府在防疫	tvbs	<a href="https://news.tvbs.com.tw/life/412042">https://news.tvbs.com.tw/life/412042</a>	15457
	15456	引領減肥風潮的新抗炎 飲食運動上場	2003-06-17	陳水扁總統6月份到蘇聯出席俄羅斯抗癌研討會，醫學大師	tvbs	<a href="https://news.tvbs.com.tw/life/411989">https://news.tvbs.com.tw/life/411989</a>	15456
	15455	清州癌症研究：以身體免疫力抗癌	2003-06-25	清州科學家最近成功研發了一種以提高身體免疫力來抑制癌	tvbs	<a href="https://news.tvbs.com.tw/life/410430">https://news.tvbs.com.tw/life/410430</a>	15455
	21346	閉月越瘦快點瘦！半年成功減重	2003-06-27	閉月越瘦快點瘦！從今年1月1日實施以來已經經過半年的評選	tvbs	<a href="https://news.tvbs.com.tw/life/410106">https://news.tvbs.com.tw/life/410106</a>	21346
	21344	減重飲食吃不停 告別煩惱身輕如羽	2003-07-01	巴西小男童奧德羅，完全不知道叔叔和外祖母為了爭取獎金	tvbs	<a href="https://news.tvbs.com.tw/life/409804">https://news.tvbs.com.tw/life/409804</a>	21344
	21342	體動到市場 與減重小孩吃午餐	2003-07-14	胖飛機大陸，國民真正瘦起來彰化市港參加兒童減重》tvbs	tvbs	<a href="https://news.tvbs.com.tw/life/408547">https://news.tvbs.com.tw/life/408547</a>	21342
	21332	胖到惹怒男友 脂肪掉1年減60公斤	2003-07-17	有一名26歲的吳小姐，遭逢家族性肥胖，體重飄到125公斤	tvbs	<a href="https://news.tvbs.com.tw/life/408130">https://news.tvbs.com.tw/life/408130</a>	21332
	15454	彰化中醫減肥成功 減脂體能抗病	2003-07-23	彰化秀傳醫院30多名中醫師，每天都會聚在一起減肥成功	tvbs	<a href="https://news.tvbs.com.tw/life/407673">https://news.tvbs.com.tw/life/407673</a>	15454
	21331	T-RUSH男生減運動 為全民減重	2003-07-23	配合第二波主打「新生代運動」，活力十足的T-RUSH特別	tvbs	<a href="https://news.tvbs.com.tw/life/407558">https://news.tvbs.com.tw/life/407558</a>	21331
	21330	局部冰敷瘦身 身體減重30公斤	2003-07-29	想要局部瘦身的女性現在有一種最新的冰敷療法，原理是利	tvbs	<a href="https://news.tvbs.com.tw/life/406876">https://news.tvbs.com.tw/life/406876</a>	21330
	21329	快量秤醫生 成功減重挑得共人體	2003-08-05	桃園市有一位重達186公斤的胖醫生吳文偉，在女網友的監	tvbs	<a href="https://news.tvbs.com.tw/life/406181">https://news.tvbs.com.tw/life/406181</a>	21329
	15453	《總統大選-民調報告》還打不響 溫尼拉蘇特拉佛羅	2003-08-07	方志華：「這裡是花蓮，剛剛炸薯的熱炸薯條，溫熱的權利	tvbs	<a href="https://news.tvbs.com.tw/topic/34/404857">https://news.tvbs.com.tw/topic/34/404857</a>	15453
	21328	轉動2個月減重60公斤的中藥方？	2003-08-19	Q：能否轉他們提供6月18號推出的新聞。就是楊小姐2個月	tvbs	<a href="https://news.tvbs.com.tw/life/404334">https://news.tvbs.com.tw/life/404334</a>	21328
	16303	瘦學溫泉現蹤 游遍天下沒裡過光	2003-08-23	您知道中國歷史上的最後一個皇帝溥儀，也喜歡泡溫泉嗎？	tvbs	<a href="https://news.tvbs.com.tw/life/403996">https://news.tvbs.com.tw/life/403996</a>	16303
	21327	阿萬仔暴瘦到300公斤 嚴重心得	2003-08-26	26歲的新竹市民呂宗惠，在消防隊員的協助下暴瘦掉體重，才	tvbs	<a href="https://news.tvbs.com.tw/life/403555">https://news.tvbs.com.tw/life/403555</a>	21327
	16302	減肥烤肉餐 訓萬人味熱量減半	2003-09-09	中秋節最熱門的活動應該非烤肉莫屬，不過對於想減肥或是	tvbs	<a href="https://news.tvbs.com.tw/life/402029">https://news.tvbs.com.tw/life/402029</a>	16302
	15451	小小耳豆按壓穴位 增強免疫力	2003-09-16	耳豆按壓穴除了可以減肥，還可以增強免疫力，保健排毒	tvbs	<a href="https://news.tvbs.com.tw/life/400886">https://news.tvbs.com.tw/life/400886</a>	15451
	21326	輕鬆胖減重45公斤 內人普遍瘦房	2003-09-24	這一點因應太胖，呼吸困難喪失生命的昌崇慶，上午在大肚	tvbs	<a href="https://news.tvbs.com.tw/life/400357">https://news.tvbs.com.tw/life/400357</a>	21326
	21325	花30萬減肥失敗 代餐減肥省下來	2003-10-01	想要快速減肥的民眾要注意了！如果一個星期減掉2公斤	tvbs	<a href="https://news.tvbs.com.tw/life/399500">https://news.tvbs.com.tw/life/399500</a>	21325
	21324	Lagerfeld04春夏 完美剪裁展性感	2003-10-10	跨洲界大師拉格斐爾，在法國發表時尚春夏的時裝，吸引洋	tvbs	<a href="https://news.tvbs.com.tw/life/398439">https://news.tvbs.com.tw/life/398439</a>	21324
	10166	怕體圓潤 200萬解剖腰帶穿著	2003-10-11	台灣每年生產大的將近三億五千萬顆腰帶，作為男性，肉肉	tvbs	<a href="https://news.tvbs.com.tw/life/398289">https://news.tvbs.com.tw/life/398289</a>	10166
	21323	85公斤小胖子減掉 一周減重2公斤	2003-11-10	亞洲盃神球熱熾，說起國內的神球熱，也有許多小胖子是	tvbs	<a href="https://news.tvbs.com.tw/life/394579">https://news.tvbs.com.tw/life/394579</a>	21323
	15450	紅色青材盒裝毒舌茶 韻化免疫力防感冒	2003-11-16	桂冠天天寒流冷，自己或是身邊的朋友是否已經有人感冒了？	tvbs	<a href="https://news.tvbs.com.tw/life/393934">https://news.tvbs.com.tw/life/393934</a>	15450

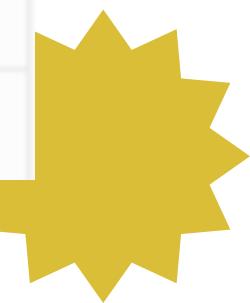


# Querying articles

(2)找出資料表health\_news和health\_news\_API中的source

```
SELECT DISTINCT source  
FROM health_news  
UNION  
SELECT DISTINCT source  
FROM "health_news_API";
```

source
"ltn"
"Yahoo Entertainment"
"Thenewslens.com"
"Gammie.com.tw"
"Jandan.net"
"Cnbeta.com.tw"
"Douban.com"
"Secretchina.com"
"Technews.tw"
"Ithome.com.tw"
"Techbang.com"
"China Times"
"Inside.com.tw"
"Cnblogs.com"
"tvba"



# Querying articles

(3) 資料以id 做升序排序

SELECT \*  
FROM health\_news  
ORDER BY id ASC;

	id	head_line	title	date	content	source	url	id
	1	男子從美國、印度返台公房 台南確診第2例		2024-10-22	出國不只慎防登革熱還要避免染上登公房！台南市這7月傳 China Times		<a href="https://www.chinatimes.com/realtimenews/1">https://www.chinatimes.com/realtimenews/1</a>	
	1	10月「國際乳癌防治月」 諸多大肆宣傳 可護健康與財富		2024-10-22	你知道粉紅絲帶象徵什麼嗎？這是乳癌防治運動的標誌。China Times		<a href="https://www.chinatimes.com/newspaper/2">https://www.chinatimes.com/newspaper/2</a>	
	1	新莊現首本土乳癌 無償快篩		2024-10-21	新北市本土乳癌首例之前分布於中和、新店、新莊19日出 China Times		<a href="https://www.chinatimes.com/newspaper/3">https://www.chinatimes.com/newspaper/3</a>	
	1	新莊出現乳癌1例！全市計67例！市府持續加強社區巡迴		2024-10-20	新北市新增1例本土乳癌病例，為新莊區40歲男，14日出 China Times		<a href="https://www.chinatimes.com/realtimenews/4">https://www.chinatimes.com/realtimenews/4</a>	
	1	乳癌疫情沒停歇！中市8人罹癌 7人仍住院接受治療		2024-10-20	近日天氣轉涼，各級病房開始添購，其中乳癌疫情沒有停歇 China Times		<a href="https://www.chinatimes.com/realtimenews/5">https://www.chinatimes.com/realtimenews/5</a>	
	1	乳癌5年存活率達9成 3招防癌發		2024-10-20	乳癌位居我國女性癌症發生率首位長達19年，且人數持續上 China Times		<a href="https://www.chinatimes.com/newspaper/6">https://www.chinatimes.com/newspaper/6</a>	
	1	乳癌隔10年仍有復發風險 專家破迷思遠3關鍵		2024-10-19	乳癌連續10年為我國女性癌症發生率首位，據統計，2022年 China Times		<a href="https://www.chinatimes.com/realtimenews/7">https://www.chinatimes.com/realtimenews/7</a>	
	1	新北本土乳癌再添1例市府加強社區巡邏		2024-10-16	新北市再添1例本土乳癌病例，為中和區41歲男，10月15日 China Times		<a href="https://www.chinatimes.com/realtimenews/8">https://www.chinatimes.com/realtimenews/8</a>	
	1	年過50血癌高警訊 醫示警：2現象恐肇癌王		2024-10-15	胰臟王有癌王之稱，家醫科醫師魏士航指出，胰臟癌虛胖才 China Times		<a href="https://www.chinatimes.com/realtimenews/9">https://www.chinatimes.com/realtimenews/9</a>	
	1	失心臟乳癌者「質抱憾夫妻第1個知道」 遺言兩極論：心寒無價		2024-10-14	女星失心臟在2021年底罹乳癌第3期，歷經了18次化療， China Times		<a href="https://www.chinatimes.com/realtimenews/10">https://www.chinatimes.com/realtimenews/10</a>	
	1	28歲女每天必嗑堅果雞胸1年養出大腸癌 醫揭3危險因子		2024-10-14	都是吃營的時候，大腸直腸外科醫師林鴻麟分享在同一天收到 China Times		<a href="https://www.chinatimes.com/realtimenews/11">https://www.chinatimes.com/realtimenews/11</a>	
	1	羅東博愛醫院 寶特瓶這樣喝		2024-10-14	宜蘭羅東博愛醫院多年推動癌症防治有成，今年在癌症防治 China Times		<a href="https://www.chinatimes.com/newspaper/12">https://www.chinatimes.com/newspaper/12</a>	
	1	羅東博愛醫院推聯癌症防治組 增佳績與高企率領先		2024-10-13	宜蘭羅東博愛醫院連續多年推動癌症防治有成，今年度在全 China Times		<a href="https://www.chinatimes.com/realtimenews/13">https://www.chinatimes.com/realtimenews/13</a>	
	1	手腳變黃以為肝病 一直竟大腸癌 2強人注意了		2024-10-12	一名40歲女子因手腳變黃，以為是肝病引發黃疸而求診，China Times		<a href="https://www.chinatimes.com/realtimenews/14">https://www.chinatimes.com/realtimenews/14</a>	
	1	「黑癌病毒」致死率近9成！疾管署升溫安撫做監戒		2024-10-11	近日「黑癌病毒」肆虐非洲撒哈拉，9月27日出現首例黑癌 China Times		<a href="https://www.chinatimes.com/realtimenews/15">https://www.chinatimes.com/realtimenews/15</a>	
	1	新北本土乳癌再添1例 6旬婦感染真菌患乳癌有驚		2024-10-10	新北市新增1例本土乳癌病例，為新店區66歲女，6日出 China Times		<a href="https://www.chinatimes.com/realtimenews/16">https://www.chinatimes.com/realtimenews/16</a>	
	1	台灣日本癌兆再添1例！安南6旬婦感染住家附近霉菌環境		2024-10-10	日本癌兆的流行高峰期已過，甚至即將進入尾聲，但台灣市 China Times		<a href="https://www.chinatimes.com/realtimenews/17">https://www.chinatimes.com/realtimenews/17</a>	
	2	抗發炎 助減重 專家推6食材：吃瓜薯CP值最高		2024-10-21	減重不只是外表，更是為了健康，營養師李婉萍表示，以下 China Times		<a href="https://www.chinatimes.com/realtimenews/18">https://www.chinatimes.com/realtimenews/18</a>	
	2	吃瘦肉燒肌肉 醫謠指食黃金時間：助燃脂更多強筋		2024-10-20	有些減肥的人，完全不敢碰黃色每碳水化合物。不過，減重 China Times		<a href="https://www.chinatimes.com/realtimenews/19">https://www.chinatimes.com/realtimenews/19</a>	
	2	「體重自賞」吃多恐傷表皮 醫示警：風險暴增40%		2024-10-20	想要減重成功，體重自賞必須要吃掉。不過，減重醫師蕭桂珍 China Times		<a href="https://www.chinatimes.com/realtimenews/20">https://www.chinatimes.com/realtimenews/20</a>	
	2	綠茶、紅茶強調不同「保護攝護腺」，醫謠最佳選擇		2024-10-19	綠茶、紅茶都是備受歡迎的茶品，到底哪種比較好？對此，China Times		<a href="https://www.chinatimes.com/realtimenews/21">https://www.chinatimes.com/realtimenews/21</a>	
	2	抹茶可防腫瘤化！8醫在功效讚：運動前喝加速燃脂		2024-10-19	抹茶是很流行的飲品，實際上它能為健康帶來許多正面影響 China Times		<a href="https://www.chinatimes.com/realtimenews/22">https://www.chinatimes.com/realtimenews/22</a>	
	2	任爾蠶蟲再破紀錄！醫謠語4秘訣 從內而外改善代謝		2024-10-19	第59屆金鐘獎昨日落幕，任爾蠶蟲憑《山裡來了鄰居》 China Times		<a href="https://www.chinatimes.com/realtimenews/23">https://www.chinatimes.com/realtimenews/23</a>	
	2	黃粉蛇討好牠不會變胖 研究：有助燃脂機制		2024-10-18	不少人都把黃粉蛇視為減重的最大敵人，但減重醫師蕭桂珍 China Times		<a href="https://www.chinatimes.com/realtimenews/24">https://www.chinatimes.com/realtimenews/24</a>	
	2	喝「通便神藥」減重沒效 醫揭關鍵：腸道發炎了		2024-10-18	一名45歲女子採用扣便不順所服，便嘗試根治盛傳的「清」 China Times		<a href="https://www.chinatimes.com/realtimenews/25">https://www.chinatimes.com/realtimenews/25</a>	
	2	早餐這樣吃減重更有效！男女代謝反應大不同		2024-10-17	本月發表在國際醫學期刊上的研究發現，男女早餐選擇影響 China Times		<a href="https://www.chinatimes.com/realtimenews/26">https://www.chinatimes.com/realtimenews/26</a>	
	2	20歲女「爆爆水水」慾望分手 醫揭關鍵原因：體白素擾		2024-10-17	一名20多歲女子，因過胖進到荷爾蒙部容易早熟，荷西光才 China Times		<a href="https://www.chinatimes.com/realtimenews/27">https://www.chinatimes.com/realtimenews/27</a>	

# Querying articles-UI

新聞公司  
China Times

主題  
stock

日期  
2024/10/09

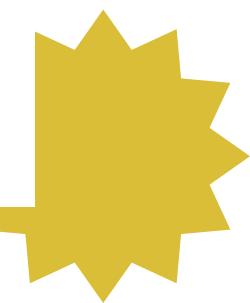
情緒  
All

開始

Title	Date	Content	Source	Emotion	URL
台中勤美洲際酒店迎賓打造全新住宿體驗	2024-10-09	【旅遊經 洪書瑱報導】根據台灣議題研究中心資訊，台中在六都觀光好感度拿下冠軍，其次為高雄、台北、新北、台南、桃園，而台中的觀光蓬勃，也吸引星級飯店及飯店品	China Times	1	<a href="#">Link</a>

topic、source、date、sentiment做索引

維護職安、母性勞工獲肯定 長榮航空領雙項殊榮	2024-10-09	長榮航空參加桃園市政府舉辦的113年桃園市推行職安衛與健康促進績優單位及人員頒獎典禮，榮獲「推行職業安全衛生優良家族」與「母性健康守護聯盟模範事業單位」雙項肯定，今日由孫嘉明總經理親自出席接受桃園市政府蘇俊賓副市長頒獎。	China Times	1	<a href="#">Link</a>
運價修正... 長榮9月營收失守500億	2024-10-09	受到國際海運市場運價修正的影響，長榮9月合併營收446.37億元，失守500億元，月營收失守500億	China Times	0	<a href="#">Link</a>



# 7. Visualization



# UI設計



左為首頁  
右為關於我們



我們提供自動化網路爬蟲，提供多來源、自適應、動態的網路爬蟲，並可以動態儲存到對應的資料庫。

我們可以動態生成對應的情緒便是分數，並生成對應的可視化圖表

詳細內容，可點選以下的github鍵以了解更詳細的資訊。

**Crawl for news and twii.**

自動化爬蟲!

Posted by Team4 on Nov 3, 2024



# 動態爬蟲 (UI)

- 選擇新聞公司  
(中時、tvbs、ltn、API)
- 選擇爬取的主題  
(Stock、Health、Sports)
- 選擇爬取的週期(頁數)

新聞公司

中國時報

主題

股市

頁數

輸入頁數

開始



# 資料索引 (UI)

- 選擇新聞公司  
(中時、tvbs、ltn、API)
  - 選擇爬取的主題  
(Stock、Health、Sports)
  - 選擇日期
  - 選擇情緒
- 以上皆可選擇不選  
(顯示該類全部)

The screenshot shows a search interface with the following fields:

- 新聞公司: China Times
- 主題: stock
- 日期: 2024/10/09
- 情緒: All

Below the filters is a "開始" (Start) button.

Title	Date	Content	Source	Emotion	URL
台中勤美 洲際酒店 10 / 23起 迎賓 打造 全新住宿 體驗	2024- 10-09	【旅遊經 洪書瑱報導】根據台灣議題研究中心資訊，台中在六都觀光好感度拿下冠軍，其次為高雄、台北、新北、台南、桃園，而台中的觀光蓬勃，也吸引星級飯店及飯店品牌進駐，目前台中的星級飯店，包括：李方艾美、林酒店、麗寶福容、台中裕元花園、台中日月千禧、台中金典、長榮桂冠、台中福華、永豐棧、台中大毅老爺等，現在民眾到台中遊玩，又有新飯店可選擇了，新開幕酒店為IHG洲際酒店集團直營管理的旗下	China Times	1	<a href="#">Link</a>

# 資料視覺化

- 選擇主題、新聞來源、日期區段
- 顯示分析、情緒分析(即時)

主題：

Health

新聞來源：

All

開始日期：

2023/01/01



結束日期：

2024/01/01



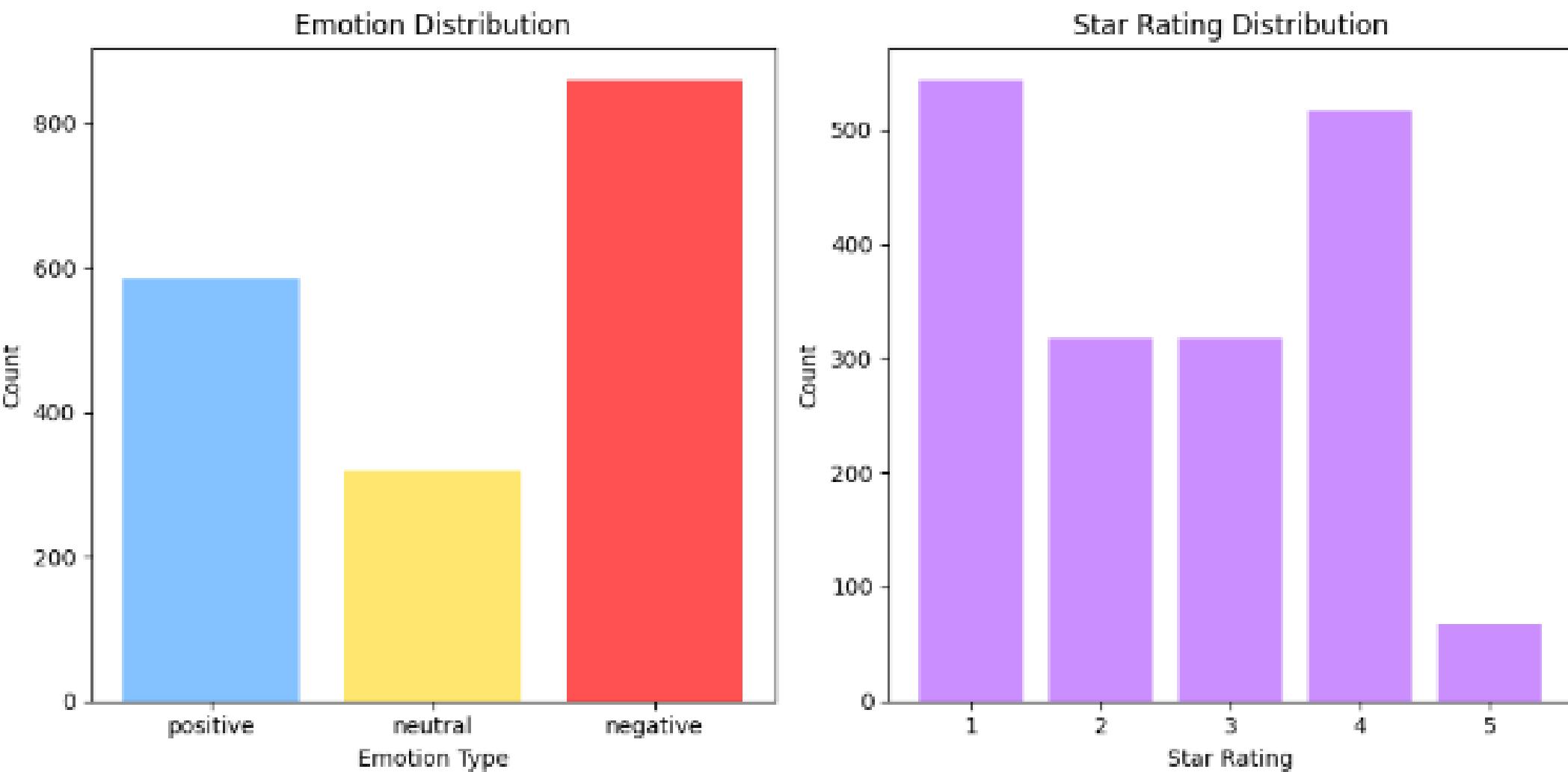
顯示分析

情緒分析

# 資料視覺化

使用 matplotlib 繪製分布圖

- 左側圖表顯示情緒類型分佈
- 右側圖表顯示星級分佈



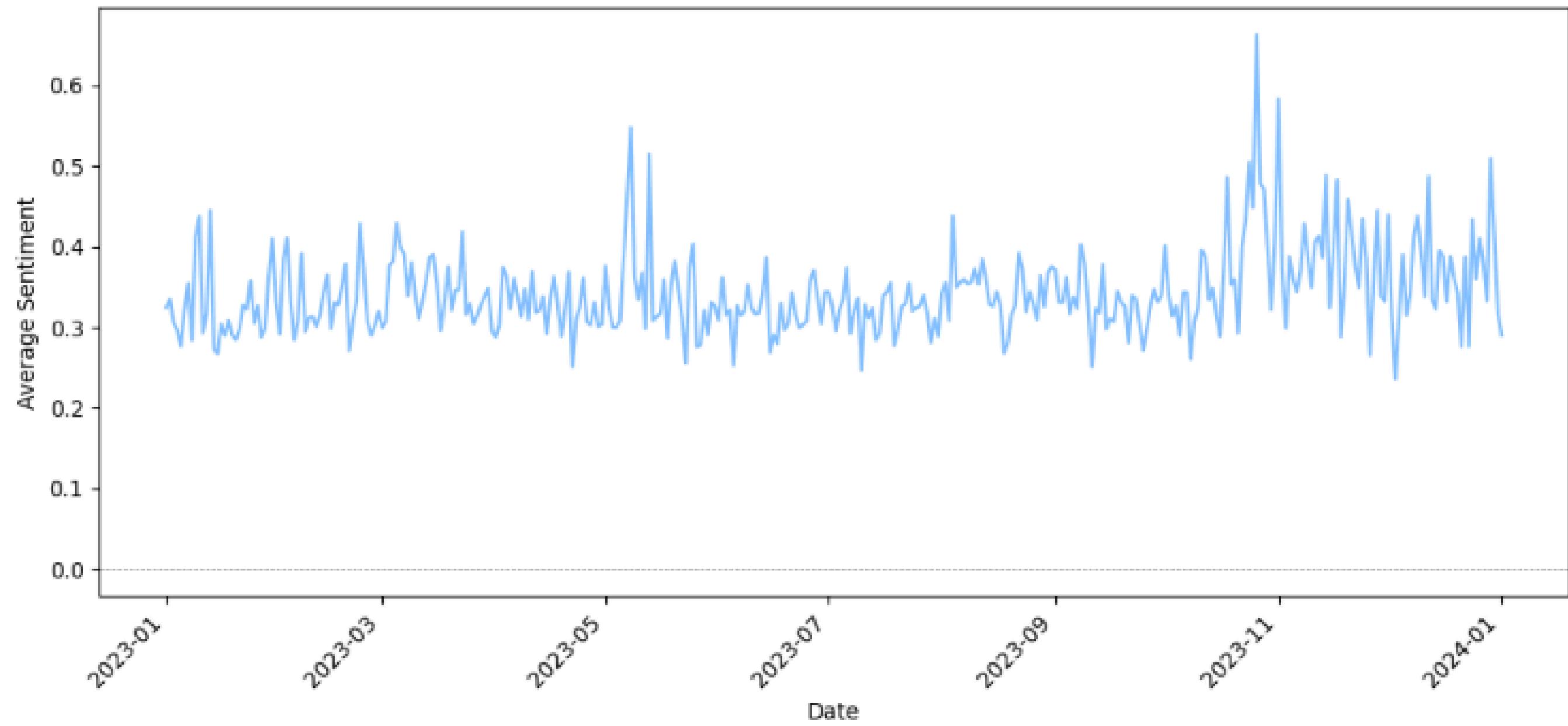
圖一為情緒(positive、neutral、negative)

圖二為star數

# 資料視覺化

繪製時間序列圖

Sentiment Trend from 2023-01-01 to 2024-01-01



# **Thank you for your listening!**

