

金融科技-文字探勘專題

文本生成-市場焦點機器人

Team2: 台大經濟四 – 洪譽家
台大財金四 – 杜宗翰
東吳巨資三 – 孫翊瑄
東吳巨資二 – 邱啟軒
台大會計二 – 蔡漢鋤

目錄

1

專案簡介

2

爬蟲與分析方法

3

網站呈現

4

討論與結語



專案簡介

解決痛點

新聞資訊龐雜

- 新聞資訊過於龐雜，使用者無法快速獲取產業或企業的相關新聞
- 無法快速得知一產業或企業近期的新聞情緒

ESG新聞需求

- 企業越發重視 ESG 投資，對於了解各企業在 ESG 等議題上著墨的需求正在上升

平台整合

- 無一客製化平台將個來源新聞摘要予以整合
- 企業內部在進行新聞議題簡報時多依靠人工蒐集整理，耗時耗力

專案發想

S&P500新聞摘要及分析



```
graph LR; A[S&P500新聞摘要及分析] --> B[焦點新聞]; A --> C[產業、企業新聞]; A --> D[ESG議題];
```

焦點新聞

產業、企業新聞

ESG議題

專案流程

新聞資料爬取

使用Goose、Newsapi爬取網路新聞平台的新聞資料與相關資訊

文字探勘、文本生成

透過Bert、textrank進行新聞的摘要、情緒分析與正負面評價

圖表、簡報呈現

利用 Streamlit 製作網頁供使用者互動及瀏覽，並自動生成簡報

組員分工

洪譽家 – 台大經濟四

- 新聞摘要與情緒分析模型
- 偏見新聞偵測
- 前端網頁建立

杜宗翰 – 台大財金四

- 關鍵字搜尋
- 條目式搜尋系統
- 前端網頁建立

邱啟軒 – 東吳巨資二

- PPT自動生成
- 成果發表影片製作
- 前端網頁建立

孫翊瑄 – 東吳巨資三

- 篩選ORG模型建立
- ESG新聞趨勢分析
- 前端網頁建立

蔡漢鋁 – 台大會計二

- 網路新聞爬蟲
- 資料庫建立
- 投影片製作

2

爬蟲與分析方法

新聞內容爬蟲



網路新聞資料

- 從網路新聞平台 Newsapi 取得新聞標題、與原始新聞連結
- 利用 Goose 從該連結爬取新聞內文



文字資料整理

- 利用套件 Pandas 整理成 DataFrame 的形式
- 再將資料依序存入三層的 Dictionary



匯出分析資料

- 將整理後的資料以 Json 的檔案格式匯出
- 將檔案上傳至雲端共享

分析方法

單篇新聞摘要 & 重要新聞選取

- Hugging Face Pytorch transformers: bert-extractive-summarizer (with BERT & KMeans)
- Gensim: TextRank

新聞情緒分析

- 財金相關詞彙情緒字典: Loughran-McDonald dictionary
- NLTK Vader

擷取新聞內提及公司

- MITIE: Named Entity Extraction、nlp stanza

自動生成簡報

- python-pptx

單篇新聞摘要、情緒分析、提及公司舉例



Genomics Market Size 2021 | Is Projected to Reach USD 94.66 Billion by 2028, Exhibiting a CAGR of 19.4%

Top Players Covered in the Genomics Market Research Report are Illumina, Inc. (California, U.S.), Thermo Fisher Scientific Inc. (Massachusetts, United States), Pacific Biosciences of California, Inc. (California, U.S.), F. Hoffmann-La Roche Ltd. (Basel, Switzerland), Agilent Technologies, Inc. (California, U.S.), Oxford Nanopore Technologies (England, U.K.), **Danaher** (Washington, D.C., U.S.), QIAGEN (Hilden, Germany), BGI (Guangdong, China), IntegraGen (Evry, France), General Electric Company (Illinois, U.S.) and other key market players

Pune, India, April 26, 2021 (GLOBE NEWSWIRE) -- The global size is projected to reach USD 94.66 billion by 2028, exhibiting a CAGR of 19.4% during the forecast period.

Fortune Business Insights™ shares this information in its report, titled “Genomics Market, 2021-2028”. As per the report, the value of the market stood at USD 23.12 billion in 2020 and USD 27.81 billion in 2021. **One of the key factors promoting the significance of this field of biology is its potential to manage zoonotic diseases, such as rabies, bird flu, plague, and the evident coronavirus.** These diseases are life-threatening and can cause huge burdens if not monitored carefully. According to the World Health Organization, in 2020, rabies accounted for nearly 55 000 deaths in Asia and Africa.

Genomic sequencing has played a vital role in identifying and diagnosing the COVID-19 virus. Continuous sequencing during the pandemic has enabled healthcare experts to closely monitor the virus, which has helped in developing effective vaccines. Such episodes have propelled the demand for genomics. Based on our findings, this market experienced robust growth of 22.6% in 2020.

Click here to get the short-term and long-term impact of COVID-19 on this market. August 2020 – Ancestry, a privately-held genealogy company launched AncestryHealth, which comprises next-generation sequencing and screening capabilities of genes associated with diseases such as colon cancer, breast cancer, blood disorders, and heart diseases. Governments worldwide are focusing on developing the healthcare infrastructure of their countries by investing in research facilities and other advancing domains in the industry. ...

偏見新聞來源舉例 -- Southwest Airlines

“ **The Points Guy (TPG)** has been publishing hands-on advice to help readers maximize their travel experiences since the site debuted in June 2010. ”

- An upgrade worth the investment: 7 thoughts from my first Southwest Airlines flight in nearly 18 months
- I just took my first flight in 453 days and it was amazing
- Cheers: Southwest will soon serve alcohol and coffee again
- Amazing deal: Fly Southwest to Hawaii for under \$300 round-trip this summer

Negative news:

- **Yahoo Entertainment:** Flight attendant loses teeth after assault. Sacramento passenger accused of attack
- **Business Insider:** A Southwest Airlines flight attendant allegedly lost two teeth after a passenger assault, and it illustrates a growing trend of unruly behavior on flights
- **USA Today:** Southwest Airlines bans passenger who 'seriously assaulted' flight attendant
- **Yahoo Entertainment:** Southwest bans woman accused of assaulting flight attendant



網站呈現

焦點新聞

×

Navigation

Latest Unbiased News

My Favorite Categories

Companies News


Keyword Search

Sentiment Analysis

ESG Media Trend

PPT Generator

Latest Unbiased News



Global Sportswear Market Outlook and Forecast Report 2021-2026: Rising Popularity of Athleisure / Increasing Participation of Women / Rise in Demand for Sustainable Sportswear

1 days ago | Yahoo Entertainment

The "Sportswear Market - Global Outlook and Forecast 2021-2026" report has been added to ResearchAndMarkets.com's offering. With the increase in participation and government initiatives, the market for women's sports and activewear will continue to grow YoY during the forecast period.

Word Cloud

Jewelry retailer Alex and Ani files for Chapter 11 bankruptcy protection

1 days ago | Reuters

Women shop for jewelry in ALEX AND ANI at the King of Prussia Mall, United States' largest retail shopping space, in King of Prussia, Pennsylvania, U.S., December 8, 2018. The company's estimated assets ranged from \$100 million to \$500 million as did its estimated liabilities, according to the filing made in the United States Bankruptcy Court for the District of Delaware.

Nike

😊

Simon Property Group Inc

😞

當週所有 S&P500 公司的
相關新聞內容文字雲

新聞連結

提及公司
新聞情緒

新聞摘要

南山人壽Team 2

關鍵字搜尋

- 使用者可輸入感興趣的關鍵字，並指定時間區間，進行新聞的搜尋
- 並在各新聞提供情緒分析的圖示與分析的數據

$$SentimentScore = \frac{(\text{正面新聞數量} \times 1 + \text{中立新聞數量} \times 0.5) \times 100}{\text{總新聞數量}}$$

新聞摘要

The screenshot shows a web application for keyword search. On the left is a 'Navigation' sidebar with buttons: 'Latest Unbiased News', 'My Favorite Categories', 'Companies News', 'Keyword Search' (highlighted with an orange arrow), 'Sentiment Analysis', 'ESG Media Trend', and 'PPT Generator'. The main area is titled 'KEYWORD SEARCH'. It features a search bar with 'Taiwan' entered, a 'GO' button, and date pickers for 'Start date' (2021/04/01) and 'End date' (2021/06/11). An orange box labeled '可選擇的時間選單' points to the date pickers. Below the search bar, a 'Sentiment Score : 53.39' is displayed in a dashed blue box. An orange arrow points from the 'Sentiment Analysis' button in the sidebar to this score. Below the score is a news article titled 'Better Buy: AMD vs. Taiwan Semiconductor Manufacturing' by 'nan | Motley Fool'. The article text discusses the semiconductor market. To the right of the article is a 'Word Cloud' section with a dashed blue box containing 'AMD', 'Xilinx', a green smiley face icon, and a plus sign. An orange box labeled '提及公司新聞情緒' points to this word cloud. Below the word cloud is another news article titled 'Taiwan chip packager King Yuan halts output after COVID cases at factory' by 'nan | Reuters'. At the bottom right, there is a button labeled 'Intel'.

產業新聞

- 從 10 種不同的產業中，選取有興趣的產業查看指定數量焦點新聞
- 設定的興趣產業在離開網站前會有記憶，不需重新設定

選取產業（可多選）

正負面圓餅圖

新聞連結

新聞摘要



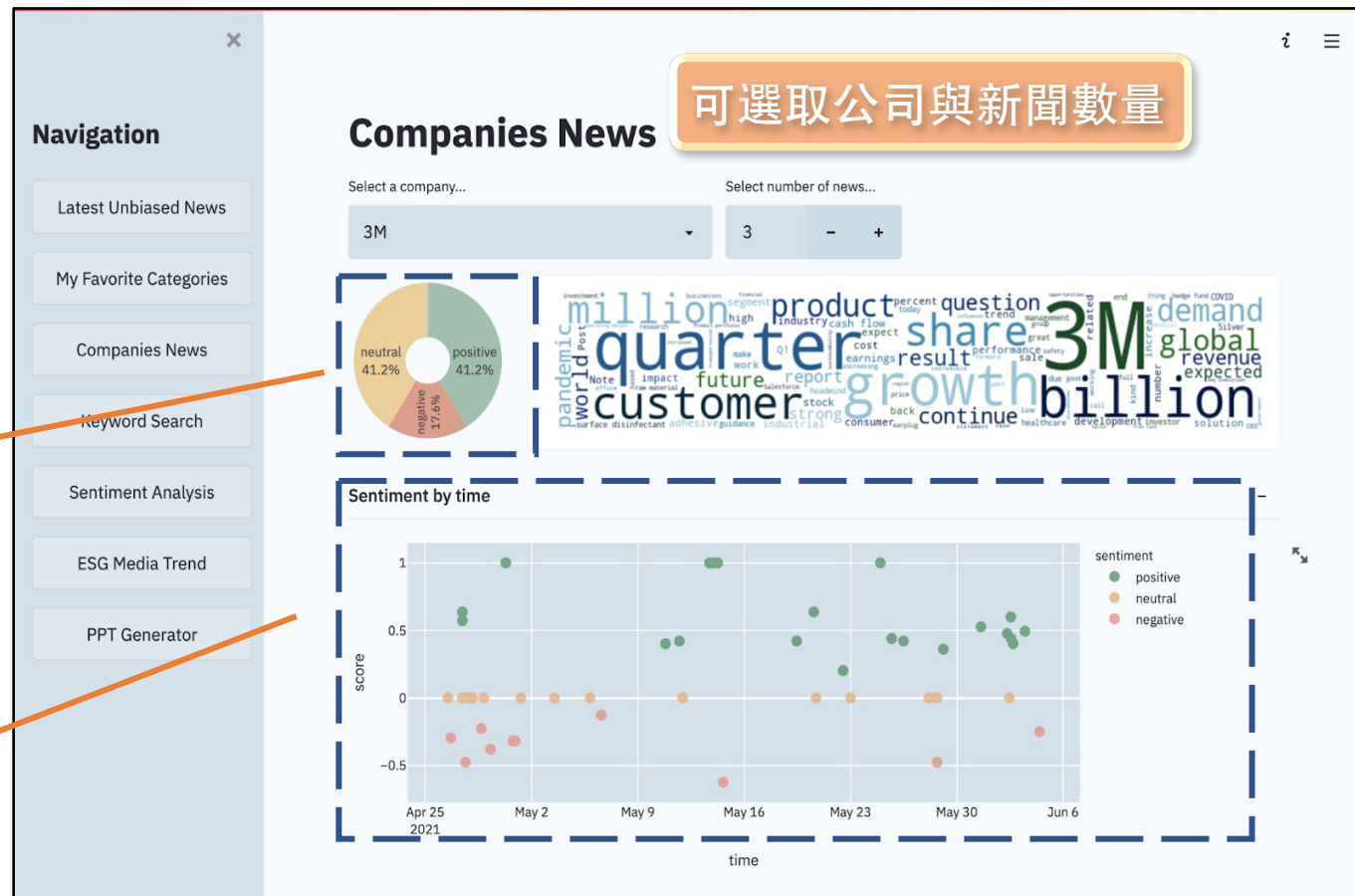
該產業新聞的文字雲

企業新聞

- 使用者可選擇感興趣的企業做搜尋
- 系統顯示相關新聞及整體情緒分析
- 並以散佈圖顯示近一個月正負面的新聞時間分佈

正負面圓餅圖

正負面新聞的時間散佈圖



公司情緒分析

- 列舉相關新聞最正面及最負面的5家企業，讓投資者快速了解企業概況

正負面圓餅圖

列舉各一篇正、負、中立新聞與其連結

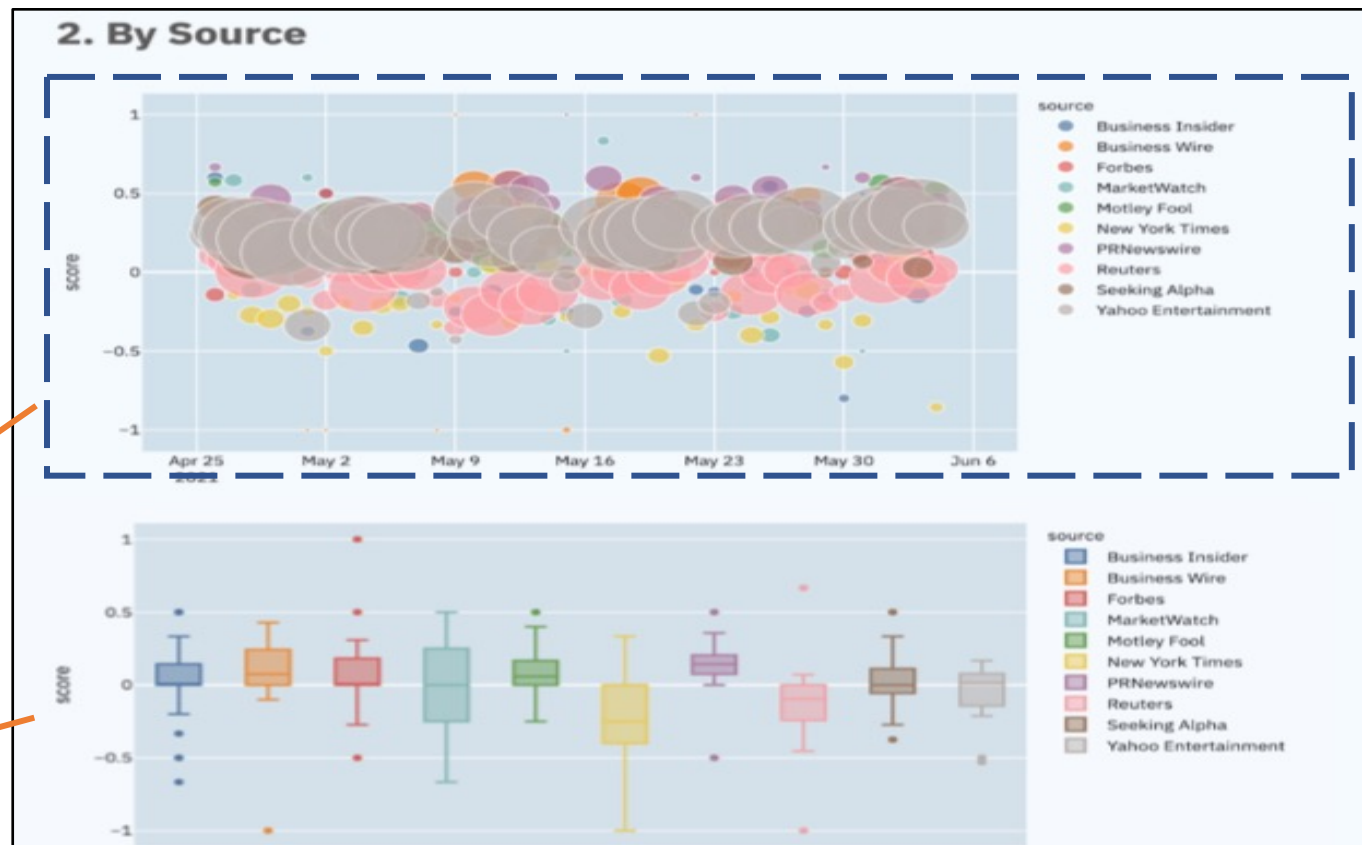


來源情緒分析

- 列舉新聞量最大的 10 間媒體，做整體新聞情緒的分析
- 讓使用者對於各媒體的整體情緒有概念，留意整體情緒落差不大的媒體 (如普遍集中報導正面新聞)

可選取新聞媒體，查看過去一個月來的新聞量及新聞情緒分數

過去一個月新聞情緒分數的箱型圖



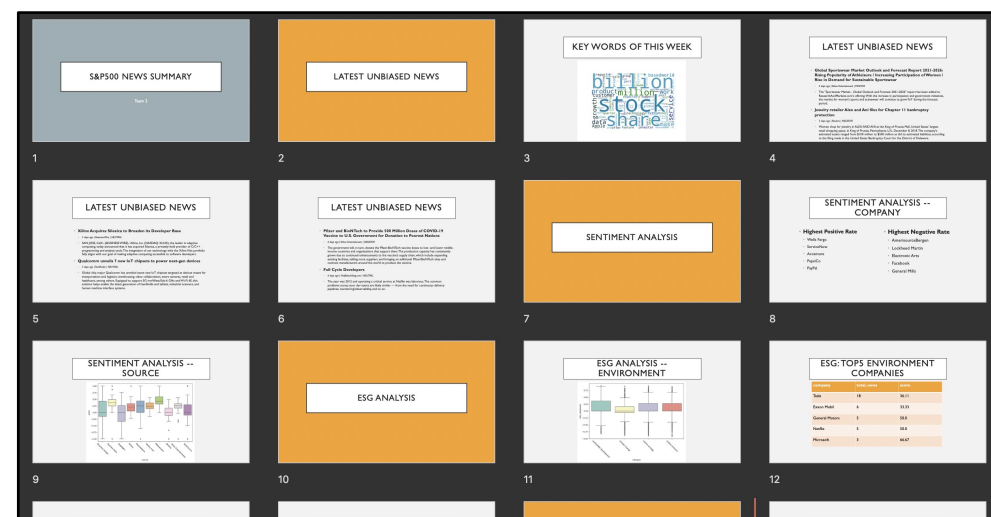
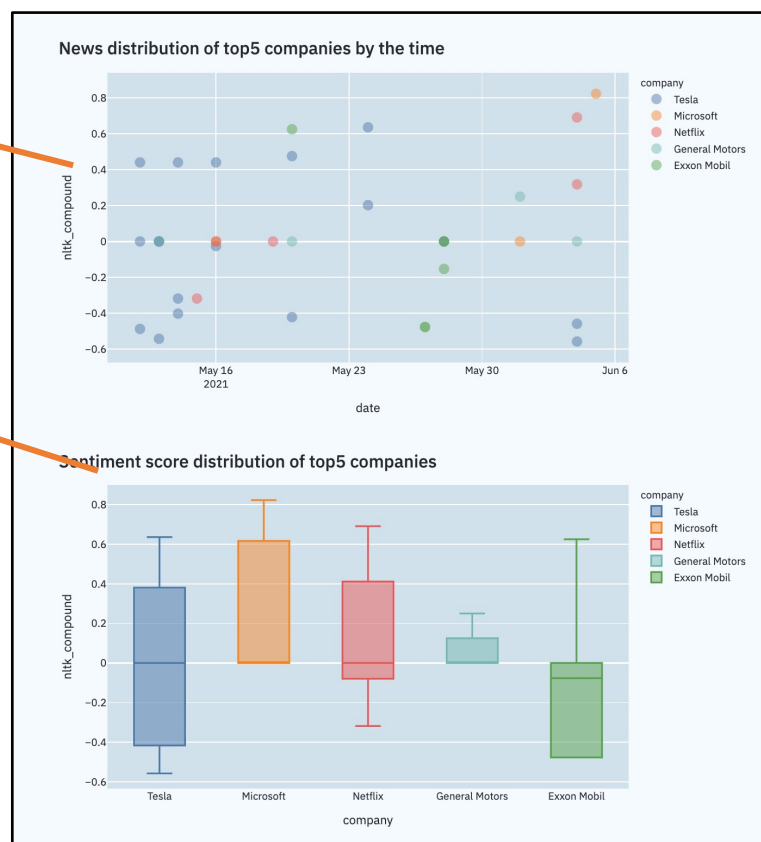
ESG新聞分析與簡報生成

- 提供環境與社會兩個主題來展示在新聞標題中的情緒分佈
- 並提供五名最常出現的組織在環境及社會相關議題表現情況

- 將上述所有內容生成簡報讓使用者匯出使用

選定議題下前五常出現組織的新聞情緒時間分佈圖

前五最常出組織的情緒分佈箱形圖





討論與結語

討論與結語

- (1) 情緒分析的精準度
- (2) 多個關鍵字搜尋 新聞量的不足
- (3) S&P500公司變動

報告結束