Finalterm Project of Public Opinion Analysis

大數據輿情分析--期末小專案報告

組員:

0624025 黃樂昀　0624041 葉秦妤

第一部分:(延續期中專案的)輿情分析任務

分析主題:

主要延續期中報告主題，分析不同領域下各個組織在網路上的討論熱度，主要分析內容有熱門關鍵詞分析、熱門組織排行分析、你的關鍵詞熱門度分析、全文檢索與關連新聞分析、你的關鍵詞情緒分析等等，且也會根據新聞類別的不同，顯示的相關組織分析結果也大不相同。

資料抓取:

三立新聞台 https://www.setn.com/（各類前 3 頁資料，11 種類別，資料比數 841 筆）

Programming Languages:

Python、Django、VScode

網頁顯示畫面:

Topic 1: 熱門關鍵詞分析：http://127.0.0.1:8000/topword/

選取想要分析的新聞類別 EX. 政治、科技、娛樂、運動……也可選擇全部，再輸入想顯示多少個關鍵字(系統預設為 10 個)，最後點選查詢按鈕，會立即顯示關鍵字繪圖與關鍵字顯示數量。

Topic 2: 熱門組織排行分析：http://127.0.0.1:8000/

選取想要分析的新聞類別，再輸入想顯示多少個關鍵字(系統預設為 30 個)，最後點選查詢按鈕，會立即顯示組織的關鍵字繪圖、關鍵字顯示數量與關鍵字雲圖。

Topic 3：你的關鍵詞熱門度分析：http://127.0.0.1:8000/userkeyword/

使用者可輸入想要分析的關鍵詞，選取條件(and/or)，再選取新聞類別與最近幾周的新聞資料，最後按下查詢，就會顯示日期時間圖表、幾篇報導提到、到的次數。

Topic 4: 全文索引與關連新聞分析

http://127.0.0.1:8000/userkeyword_assoc/

使用者可輸入想要分析的關鍵詞，選取條件(and/or)，再選取新聞類別與最近幾周的新聞資料，最後按下查詢，會立即顯示哪些關鍵字與它同時出現、有關的新聞連結、關鍵字出現的段落與它同時出現的關鍵字顯示數量。

以下新聞與它有關

- 科技:羨慕！台積電分紅平均每人領139萬
- 科技:下一個護國神山在哪？張忠謀親吐1字
- 生活:台灣啥東西最強？網曝這項：獨步全球
- 生活:羨慕！台積電分紅平均每人領139萬
- 財經:亞洲金融雜誌：台積電最佳管理公司
- 財經:羨慕！台積電分紅平均每人領139萬
- 財經:電金傳產並進　台股收17572點
- 財經:受惠晶片荒　聯電股價創20年新高
- 財經:聯發科新天價1090元　登台股第2
- 財經:中國半導體遇困境　法學者：台灣優勢

First « 1 2 3 4 » Last Page 1/4

關鍵字所在的段落

- 記者谷庭 / 台北報導【更新內容：平均每位員工可領139萬元】台積電（2330）、世界先進（5347）將進行員工年度調薪，台積電所有員工平均加薪幅度3%到5%，世界先進則平均幅度約3%，將在月底前實施。
- 而台積電也將分紅，平均每位員工可領139萬元，煞羨旁人。
- ▲台積電將分紅，平均每人可領得139萬元。
- （圖 / 資料照）台積電今年1月開始調薪，底薪調薪幅度約達20%，也是有史以來加薪幅度最大的一次，而每年4月也會再進行年度調薪，台積電表示，所有基層員工到主管等，平均加薪幅度約3%到5%，依績效表現而定。
- 除了加薪幅度大，公司獲利員工分紅更是驚人，像台積電今年為例，包括業績獎金與酬勞及分紅新台幣695.11億元，平均每人可領得139萬元，較去年平均數增約36萬元。
- 記者陳政宇 / 台北報導全球晶圓代工龍頭台積電被視為我國的「護國神山」，不過，台灣能否再造一座護國神山。
- 台積電創辦人張忠謀今（21）日出席「2021大師智庫論壇」時回應直言，「難。
- （圖 / 記者陳弋攝影）不過，台灣能否複製台積電模式，再造下一座「護國神山」。

First « 1 2 3 4 5 » Last Page 1/5

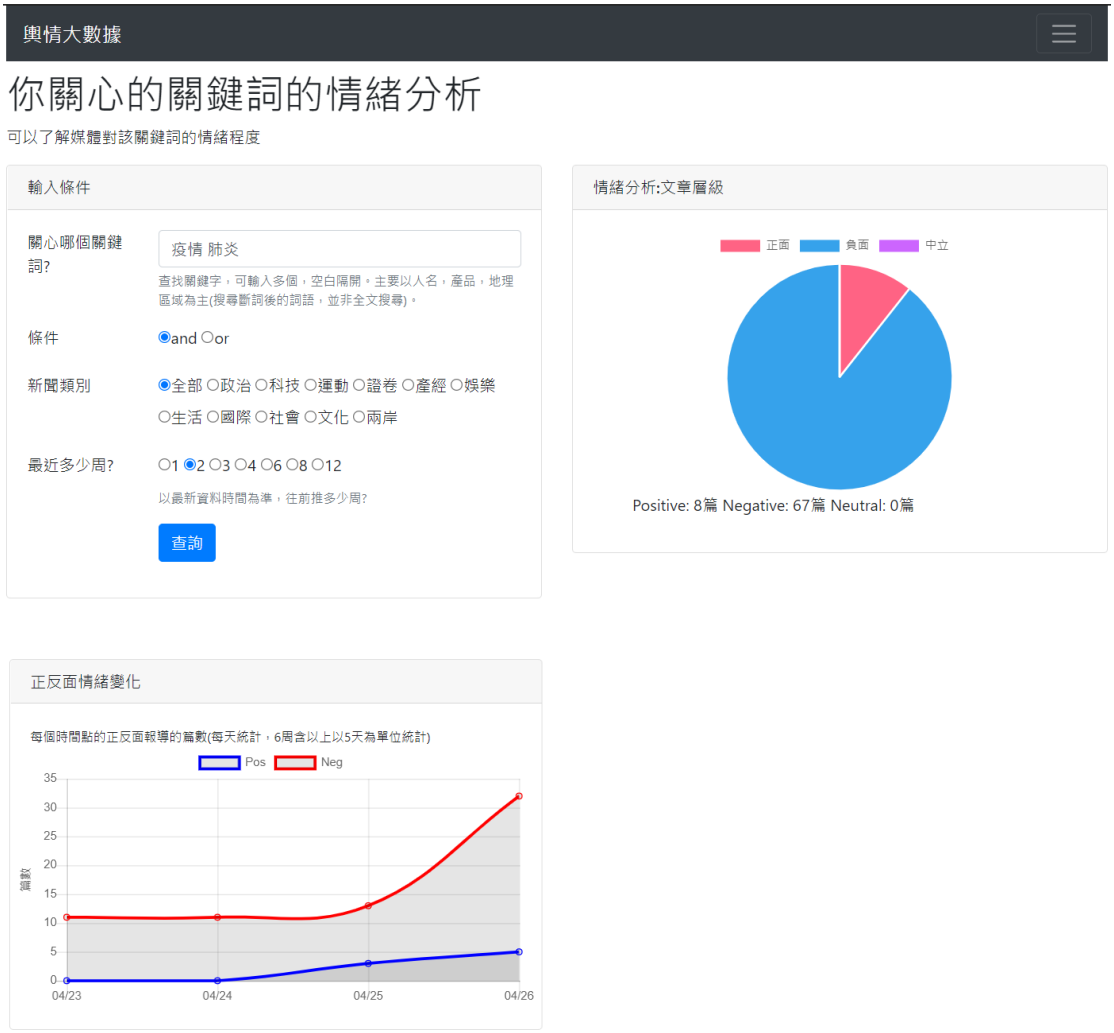
與它同時出現的關鍵字

- 台積電,84
- 台灣,63
- 經濟,28
- 台股,28
- 員工,24
- 漲幅,22
- 投資,22
- 幅度,21
- 指數,21
- 半導體,21
- 世界,19
- 聯電,18

Topic 5: 你的關鍵詞情緒分析：

http://127.0.0.1:8000/userkeyword_sentiment/

使用者可了解媒體對於該關鍵詞的情緒程度，輸入想要分析的關鍵詞，選取條件(and/or)，再選取新聞類別與最近幾周的新聞資料，最後按下查詢，會立即顯示情緒分析圓餅圖與不同時間正反面的情緒變化。



第二部分: 深度學習任務

主題:真假新聞辨別

在這個真假新聞參雜的世代，人們面對網路上的資訊無法確切的分辨其內容的真偽，所以我們希望能夠訓練一個模型，可以自動找出假新聞以節省人工檢查的成本。資料集則是由中國的手機新聞應用：今日頭條的母公司字節跳動所提出的。（知名的抖音也是由該公司的產品）

使用的資料集(Kaggle 上公開的資料集)

WSDM-Fake News Classification

https://www.kaggle.com/c/fake-news-pair-classification-challenge

訓練資料集（Training Set）約有 32 萬筆數據、測試資料集（Test Set）則約為 8 萬筆

參考資料:

進入 NLP 世界的最佳橋樑:寫給所有人的自然語言處理與深度學習入門指南

https://leemeng.tw/shortest-path-to-the-nlp-world-a-gentle-guide-of-natural-language-processing-and-deep-learning-for-everyone.html#top

訓練資料集一部分的內容如下:

	A title1_zh the fake news title 1 in Chinese	A title2_zh the news title 2 in Chinese	A title1_en the fake news title 1 in English	A title2_en the news title 2 in English	A label indicates the relation between the news pair: agreed/disagreed/unrelated
	69170 unique values	138434 unique values	67869 unique values	136111 unique values	unrelated 68% agreed 29% Other (1) 3%
14	"用大蒜鉴别地沟油的方法，怎么鉴别地沟油	翻炒大蒜可鉴别地沟油	"How to discriminate oil from gutter oil by means of garlic.	stir-fried garlic to identify gutter oil	agreed
15	"飞机就要起飞，一个男人在机舱口跪下！"这是今天最催泪的一幕……	陈乔恩公开宣布与他分手：有时候该放手就不再留恋	"The plane is about to take off. A man knelt down at the hatch." That was the tear-jerking scene today...	Chen Qihan publicly announced to break up with him: sometimes when you have to let go	unrelated
16	"飞机就要起飞，一个男人在机舱口跪下！"这是最催泪的一幕	「网警辟谣」飞机起飞前男人机舱口跪下？这故事居然是编的！	"The plane is about to take off. A man knelt down at the hatch!" That was the tear-jerking scene.	The man on the front of the cockpit kneeling before takeoff? This story is made up!	disagreed
17	"男人在机舱口跪下！"原来一切都只因为爱！	"父亲跪舱门口"谣言实锤！微博首发者：上了假冒飞行员的当！	"Men kneel at the hatch!" All because of love!	"Father kneel door" rumor mill! Microblog starter: cheated by the fake pilots!	unrelated

*第一欄位 title1_zh 代表的是「已知假新聞」 A 的中文標題：

用大蒜鉴别地沟油的方法,怎么鉴别地沟油

*第二欄位 title2_zh 則是一筆新的新聞 B 的新聞標題還未確定真偽

翻炒大蒜可鉴别地沟油

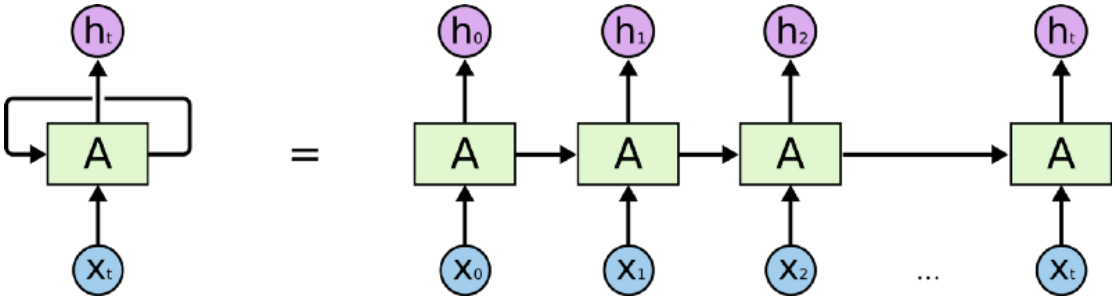
訓練方式:

如果新聞 B 同意假新聞 A 的敘述的話，我們可以將 B 也視為一個假新聞；而如果 B 不同意假新聞 A 的敘述的話，我們可以放心地將 B 新聞釋出給一般大眾查看；如果 B 與 A 無關的話，可以考慮再進一步處理 B。

循環神經網路（Recurrent Neural Network, 簡稱 RNN）

有記憶的循環神經網路

RNN 是一種有「記憶力」的神經網路，其最為人所知的形式如下：



RNN 非常適合拿來處理像是自然語言這種序列數據，就像一般人閱讀一樣，是由左到右逐字在大腦裡處理這些文字，同時不斷地更新腦中的記憶狀態。

每當下個詞彙映入眼中，腦中的處理都會跟以下兩者相關：

- 前面所有已讀的詞彙
- 目前腦中的記憶狀態

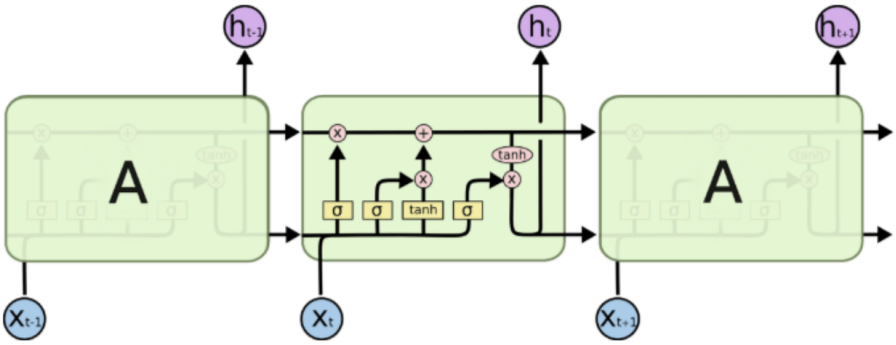
利用內在迴圈以及細胞內的「記憶狀態」來處理序列資料。

但 RNN 有一個缺點，它沒辦法很好地「記住」前面處理過的序列元素，造成 RNN 在處理後來的元素時，就已經把前面重要的資訊給忘記了。就好像一個人在講了好長一段話以後，忘了前面到底講過些什麼的情境。

因此本次訓練是以下面這個方式進行…

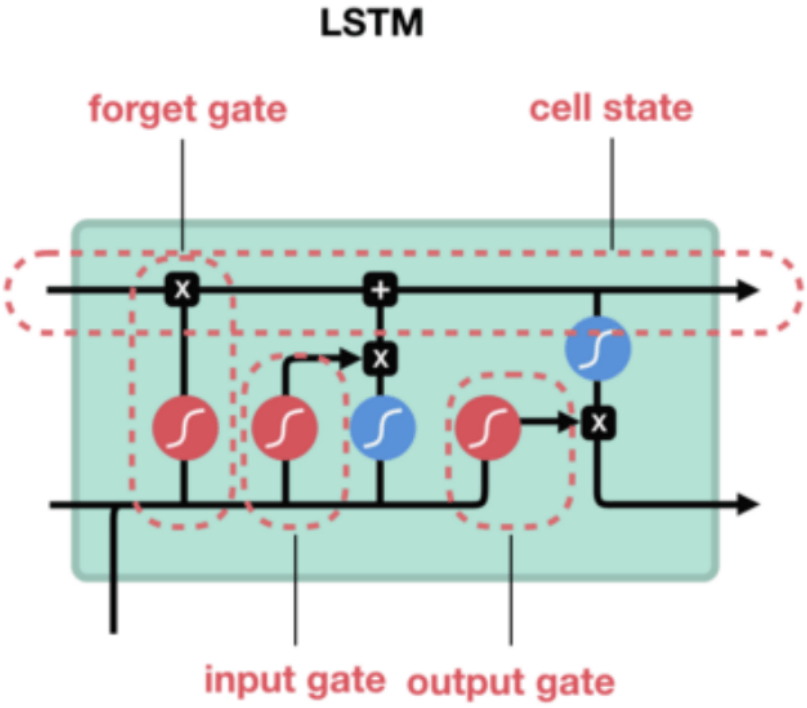
長短期記憶（Long Short-Term Memory，後簡稱 LSTM）

記憶力好的 LSTM 細胞



基本上一個 LSTM 細胞裡頭會有 3 個閘門（Gates）來控制細胞在不同時間點的記憶狀態：

- Forget Gate：決定細胞是否要遺忘目前的記憶狀態
- Input Gate：決定目前輸入有沒有重要到值得處理
- Output Gate：決定更新後的記憶狀態有多少要輸出



因此 LSTM 可以將很久以前的記憶狀態儲存下來，需要的時候再拿出來使用