網路新聞分析與推薦

資料探勘與文本分析期末報告

第四組

資三B 08156202 王泳泠

資三B 08156212 劉紫暄

資三B 08156244 廖育琳

目錄

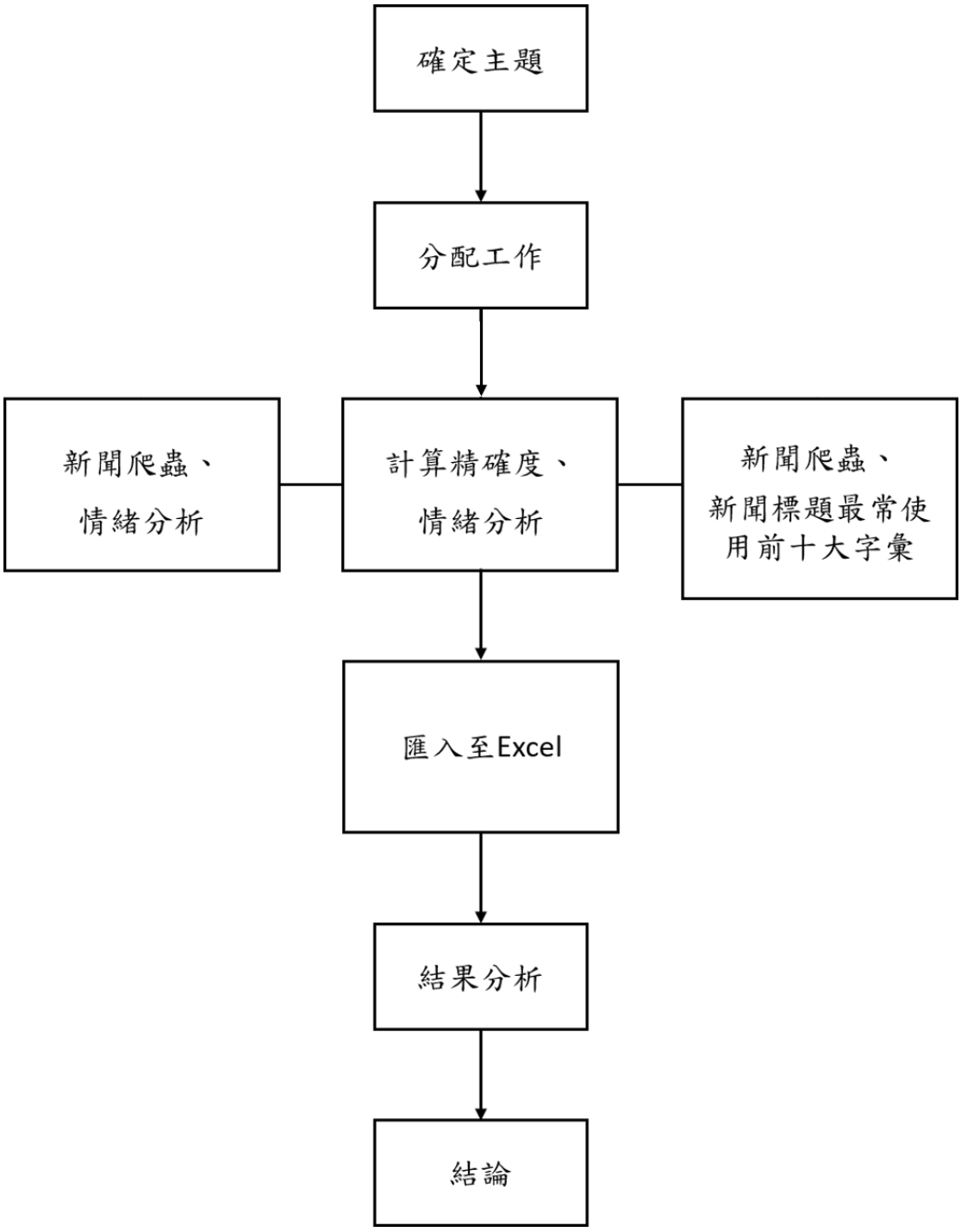
一、 研究動機與目的

隨著網路經濟體系越來越成熟，人們透過身邊的任何智慧型裝置便能隨時隨地查看新聞，了解世界當下發生了甚麼事，但當你點擊了一篇非常有噱頭的新聞，卻發現內文卻不像標題所說的那麼誇張，甚至還有內文與標題毫不相關的新聞，每次遇到此情況，一早起床好奇世界變化的好心情，難免會受到影響與感到失望。

因此我們以三個台灣網路流量前三新聞媒體中熱門板新聞來做比較與分析，之所以採用熱門板，是因為有許多民眾都對該篇新聞感興趣時，該篇新聞才會出現在熱門板，因此該板代表民眾會受吸引的新聞標題，透過研究分析該板標題，可以得知是否有使用誇大用詞來吸引讀者點擊，造成標題與內文不相符等情況，但採用熱門板也會造成一些問題，因為熱門板經常隨著該篇新聞觀看人次而有所變動，因此單爬取一次網頁作分析會造成分析內容過於限定，為了解決此問題，我們將多天、多次爬取各個不同時段的新聞取其平均值。

將標題及內文使用 CKIP 技術進行中文斷詞，接下來，採取餘弦相似度(Cosine Similarity)計算方式，判斷標題與內文所述內容是否相符合及利用 SnowNLP 技術分析內文每個句子帶來的平均情緒值，透過以上分析結果來推薦讀者在特定情況下可閱讀何者新聞網站較為適合。

二、 研究流程


確定主題
分配工作
新聞爬蟲、
情緒分析
計算精確度、
情緒分析
新聞爬蟲、
新聞標題最常使
用前十大字彙
匯入至Excel
結果分析
結論

三、研究內容

3.1 網頁爬蟲

我們以台灣網路流量前三名的網站(ETtoday、自由時報、三立新聞網)之熱門板新聞作為爬蟲目標，分別在每日三個時段(早、中、晚)爬取熱門新聞。(以下內容將會以爬取 ETtoday 為例)

ETtoday 爬蟲程式碼:

```
title = []
urls = []

# 熱門版
u = "https://www.ettoday.net/news/hot-news.htm"
res = requests.get(u)
soup = BeautifulSoup(res.content, "lxml")
soup = soup.find_all("div", class_="piece clearfix")

domian = "https://www.ettoday.net"
for i in range(len(soup)) :
    if len(urls) < 50 :
        url = "https://www.ettoday.net" + soup[i].select("a")[0]["href"]
        urls.append(url)
        t = soup[i].select_one("h3").text
        title.append(t)

# 抓內文
allcontent = []
for u in urls:
    content = []
    res = requests.get(u)
    soup = BeautifulSoup(res.content, "lxml")
    try:
        soup = soup.find("div", class_="story")
        #print(soup)
        for a in soup.find_all("p"):
            p = a.text
            check = 0
            for f in range(0, len(p)-1) :
                if p[f:f+2] == '圖/' or p[f] == '▲' or p[f] == '►' or p[f] == '▼':
                    check = 1
            if check != 1 and len(p) > 1:
                content.append(p)
        content_str = ''
        for i in range(1, len(content)) :
            content_str = content_str + ' ' + content[i]
        allcontent.append(content_str)

        time.sleep(2)
    except:
        pass
if len(urls) == len(title) == len(allcontent) :
    print('OK')
```

```
OK
```

爬蟲完畢且無問題將會回傳 OK，讓我們知道已完成 ETtoday 在該時段熱門新聞的爬蟲。

爬蟲結果:

```
# 網址 + 標題 + 內文
for i in range(len(allcontent)) :
    print(urls[i] + '\n' + title[i] + '\n' + allcontent[i])
    print('-----------------------------------------------------------------------------------------------------------------------')
```

```
https://www.ettoday.net/news/20220611/2270618.htm
快訊／趙薇消失9個月突悲痛發文：我一無所有！ 116字曝最新近況
 大陸女星趙薇於2021年8月26日無預警所有作品被下架，粉絲超話被關閉，被點名為劣跡藝人，外傳「被消失」，網路上甚至流傳著一份「25人封殺名單」，她的名字就被排在第一位，讓不少網友擔憂她的近況，事隔近1年，她總算發文，卻難掩悲傷：「一片落葉，一朵飛花，一絲輕煙！」 趙薇10日在IG限時動態發文悼念過世爸爸，「生生不息！阿彌陀佛，爸爸永遠在我心裏，如一片落葉，一朵飛花，一絲輕煙！這些人類可想像的描述都無比蒼白！如野草，生生不息，在我有限的認知裏，不會離開我，就像我也不會離開你。至於離開是什麼？根本沒有離開，你的一切不值一提，而我一無所有。」 這也是趙薇消失9個月以來的最新發文，她PO出爸爸的照片寫道：「爸爸，你就像野草一樣對嗎？生生不息，生生世世利益一切有情眾生，離苦得樂，證悟成佛！南無阿彌陀佛！」雖然並未提到自己目前行蹤，但言語之間流露出淡然的意味，而後她又將貼文刪除，可說是相當低調。 趙薇自遭到一夜清算後，目前仍下落不明，遭消失的原因也沒有答案，但外傳她可能捲入先前爆發的杭州市委書記周江勇貪腐案，以及和馬雲遭清算有關，有消息指出，大陸中央正在調查一件未公開的重大案件，疑似就是她被消失的真正原因。 趙薇後續行蹤成謎，圈內好友也與她失聯，微博上瘋傳她飛至法國鄉下酒莊與老公會合，但都沒有官方證實，事後現身社群網站發文，表示自己目前人與爸媽待在一起，還在回覆網友的留言中暗示自己「人還在北京」，後續也有網友爆料她「逃不了」。
-----------------------------------------------------------------------------------------------------------------------
https://www.ettoday.net/news/20220611/2270726.htm
快訊／今本土+79598例 莊人祥曝數字大增原因：轉檔異常
 中央流行疫情指揮中心今（11）日公布新增7萬9598例本土個案；另增65例境外移入，確診個案新增211例死亡。指揮中心表示，由於電腦程式一度出現轉檔異常，讓昨晚數據未進入系統，今早經修復後再次統計，病例數納入早上新通報個案，也因此病例數比昨天要明顯增加。 今天指揮中心記者會最開始並未公布新增病例總數，指揮中心發言人莊人祥說明，由於早上電腦倉儲程式有些問題，還在整理當中，稍後有資料會再說明；後續於記者會開始約10分鐘後宣布，今日新增病例總數為7萬9663例，當中有7萬9598例本土個案；65例境外移入。 今日本土病例比昨天明顯要多，莊人祥解釋，每天單日新增病例統計結算是到當日深夜零時為止，然而昨天晚上10點後系統「轉檔異常」，後續新增個案沒有轉入，今早統整數據時發現部分縣市明明有上傳資料，但法傳系統卻
```

3.2 新聞標題分析

能讓我們快速了解一篇新聞內容的是新聞標題，若標題不吸引人，就沒有人會點進去瀏覽內容。在新聞業如此競爭激烈的時代，每天新聞上百則，又有很多新聞台、報紙。怎麼樣的標題才會吸引人點進去看呢？因此，我們先究標題分析出標題前十大字彙（檢視出現最多的十個「字」是哪些），並以 tf-idf 加權技術萃取各家新聞的關鍵字，並且以文字雲作為視覺化描述。

3.2.1 標題前十大字彙

計算在所有熱門新聞標題中出現最多次數的字。

程式碼：

```
# 去除標點符號+數字+英文字母的標題&文章
titles = str(title)
titles1 = ''.join(char for char in titles if char.isalnum())
titles2 = ''.join([i for i in titles1 if not i.isdigit()])
titles3 = re.sub('[a-zA-Z]','',titles2)

# 每個字出現次數
wordt = {}
for w in titles3:
    if w not in wordt:
        wordt[w] = 1
    else :
        wordt[w] = wordt[w] + 1
```

3.2.2 標題文字雲

使用 jieba 套件，將所有新聞標題斷詞後，使用 TF-IDF 算法依照詞頻權重排列，並製成文字雲，快速識別標題中權重高的詞彙。

程式碼：

```
titles1 = ''.join(char for char in str(title) if char.isalnum())
titles2 = ''.join([i for i in titles1 if not i.isdigit()])
titles = re.sub('[a-zA-Z]','',titles2)

tags = jieba.analyse.extract_tags(titles, topK=50, withWeight=True)
dictionary = {}
for i in range(len(tags)) :
    dictionary[tags[i][0]] = tags[i][1]

# 設定文字雲細項
color_list=['#000000','#A9A9A9','#696969']# 建立顏色數組，更改字體顏色
colormap=colors.ListedColormap(color_list)
fontpath = "C:/test-wordcloud/msjh.ttc"  # 字型檔
mask = np.array(Image.open("C:/mickey.png")) # 文字雲樣式
wordcloud = WordCloud(background_color="white", colormap=colormap, mask=mask, font_path=fontpath)
wc = wordcloud.generate_from_frequencies(dictionary)

plt.figure(figsize=(10,10))
plt.imshow(wc), plt.axis("off")
plt.show()

# 會存在下載中，注意檔名重複問題
wc.to_file('et_wordcloud.jpg')
```

標題十大字彙結果：

	ETTODAY			自由時報			三立新聞網		
	精確度	情緒度	十大字彙	精確度	情緒度	十大字彙	精確度	情緒度	十大字彙
06/09 16:50	63.25%	41.94%	人曝光一年萬女死房中	63.83%	47.97%	死網手山單台分達女報	63.77%	49.78%	字一網國了台女小大南
06/09 23:51	63.29%	41.18%	一月網光不女曝萬神了	60.75%	47.52%	大南車死公手台長不重	61.96%	48.63%	網皇了小國人來大家字
06/10 11:46	65.22%	41.53%	女一月光手大了變人曝	59.59%	43.90%	中不新出台長光公事斯	62.46%	50.38%	不國人蔣大介石這歲中
06/10 17:34	66.79%	40.96%	女人曝月歲大光公開一	60.16%	43.24%	中出不大台賴人新球有	60.17%	49.98%	蔣國不介石女字人這有
06/10 23:12	67.68%	40.14%	女快訊高一車曝光後國	62.02%	45.47%	中不大賴出一台了球高	62.10%	48.45%	蔣介石最人不後曝歲這
06/11 10:19	65.24%	40.51%	屍前一被車了高不夫妻	63.58%	42.83%	中一大國下不出天他會	61.02%	48.76%	人女大最國蔣介石不歲
06/11 14:10	64.69%	40.77%	屍一前不女曝了快訊命	65.61%	41.77%	中國一不人無了這他下	61.92%	45.91%	人大女曝國不光名都一
06/11 22:28	63.16%	40.78%	曝不女了快訊一光人大	66.03%	42.02%	中一台不點國有好重灣	64.75%	46.37%	人不大女國名最石中曝
06/12 10:05	64.97%	39.47%	不一人年台男曝被老公	56.69%	45.23%	不中了倒回台車週下一	63.19%	46.32%	不一人曝見女字看石這
06/12 14:34	63.35%	41.08%	不曝年大一了我台中女	62.04%	46.04%	不中大了台本賽市週一	61.10%	47.25%	中女國人是有台新不本
06/12 20:10	61.93%	40.79%	不一人中台了女我老快	63.10%	42.68%	中國不台大賽市防擊看	60.79%	48.06%	人一不中台有女國網了
06/13 11:00	63.52%	42.41%	女不年了一大人到星是	68.27%	40.09%	黨不賽倫市週新戰印尼	61.38%	47.75%	中爆看不了後照一網字
06/13 14:14	63.15%	42.16%	大了年女不星到一車是	66.98%	41.42%	不黨天市新黃文戲戰人	60.70%	48.37%	中了大曝爆人女後不字
06/13 20:10	63.16%	42.38%	了大一年女到曝人不是	63.96%	41.37%	不新人一黃文戲市下看	61.29%	48.99%	了女大人曝出後天送不
06/14 10:09	63.41%	43.98%	一人曝被內年光最家出	60.14%	47.94%	國了軍俄新台人年中美	60.00%	48.23%	大他了被天爆女人字取
06/14 14:00	64.44%	43.83%	人曝一被不女內出年真	60.09%	47.81%	國軍中人大台美入新俄	62.43%	48.48%	女了上大他爆被名不曝

標題文字雲結果：

ETToday　　自由時報　　三立新聞網

由於標題以較少字數概要的說明新聞內容，更需精簡選擇用字，不過在標題中時常可以看見「死」、「曝」、「屍」等字。除此之外，經過 TF-IDF 加權技術萃取關鍵字，並且製成文字雲後，可以看出新聞標題包含「重症」、「確診」、「悲劇」等。因此我們推論這些字彙和詞語容易吸引讀者的注意，而新聞，雖然現今新聞寫稿越來越嚴謹，但許多新聞為了吸引民眾的點閱來提高流量，使用較為誇張、聳動的字或詞。在這些用字和用詞下，乍看下讓你很想點進去看的標題，點進去後卻發現文不對題。時間一久，讀者會對該媒體產生不信任感。

因此，我們使用文字探勘技術，以精確度來了解台灣網路媒體前三名新聞網標題用字和用詞與內文是否相符，另外，也透過情緒分析新聞網的立場是否中立，並依照分析出的結果來提供民眾選擇傾向的新聞風格。

3.3 平均精確度---餘弦相似、相符合程度

在判斷標題與內文相似度部分，我們首先採用 CKIP 斷詞方式分別把標題及內文斷詞，採取餘弦相似度(Cosine Similarity)計算方式，來判斷標題與內文所述內容是否相符合。

- 標題與內文斷詞 ：

將標題及內文去掉不用進行斷詞的字詞(例如:標點符號及英文單字)，方便斷詞後計算精確度與情緒度。

斷詞技術我們採用的是 CKIP，CKIP 為中研院資訊所、語言所於民國 75 年成立的中文語言小組所開發，在繁體中文的自然語言處理中，CKIP 是斷詞最精確的工具。

程式碼如下 ：

```
# 刪除標點符號、數字及英文
stop_word01 = ['「', '」',' ' , '（' , '）' , '/'  , '：'  , '-' ,'，','。','\n','\n','！','\u3000', '？', '《', '》', '.']
def clearData(news , stop_words):
    result = ''
    re1 = re.sub('[a-zA-Z]','',news)
    for w in re1 :
        if w not in stop_words :
            result = result + w
    return result

# 內文
token_c=[]
for i in allcontent:
    token_c.append(clearData(i,stop_word01))

# 標題
token_t=[]
for i in title:
    token_t.append(clearData(i,stop_word01))
```

```
# CKIP斷詞
from ckiptagger import WS
ws = WS("./data")
ws_title=[]
ws_content=[]
for i in range(len(token_t)) :
    title_text = token_t[i]
    content_text = token_c[i]

    ws_title.append(ws([title_text]))
    ws_content.append(ws([content_text]))
```

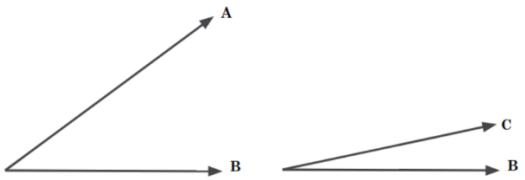
斷詞後，使用餘弦相似度方法來分析標題內文相似度

- 餘弦相似度 Cosine Similarity

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

公式:

Cosine Similarity（餘弦相似度）是在計算文本相似度時相當常見的一種計算方法，原理也相當易懂，基本上就是計算『兩向量』之間的 Cosine 夾角。夾角越大(Cosine 值越接近 0)代表兩個向量越是不像；夾角越小(Cosine 值越接近 1)，代表兩個向量越是相像。



以上圖來說，表示 C 向量與 B 向量最為相似。

運用此公式，我們透過斷詞完的各個標題與內文，將出現在標題的字詞是否出現在內文，並計算他出現在內文次數，把標題字詞出現次數轉換成向量，因為以標題字詞次數向量代表上圖的 B 向量，故每個標題字詞次數皆為 1。

程式碼如下:

```
veclist_title=[]      # 標題內文向量存取陣列
veclist_content=[]

for a in range(len(ws_title)):
    countword={}      #統計標題斷詞在內文斷詞中出現次數
    for i in ws_title[a][0]:
        countword[i]=0
        for j in ws_content[a][0]:
            if i.__eq__(j):
                if i not in countword:
                        countword[i] = 1
                else:
                        countword[i] +=1

    vector_content=[]
    vector_title=[]
    for k,v in countword.items():
        vector_content.append(countword.get(k))      # content 次數向量
        vector_title.append(1) # title 次數向量 (因以標題為主要判斷出現次數之文字依據,故向量值皆為1

    veclist_content.append(vector_content)  # 加入標題內文向量存取陣列
    veclist_title.append(vector_title)

```

```
# 向量輸出範例
print("Title vector example : {}".format(veclist_title[3]))
print("Content vector example : {}".format(veclist_content[3]))

```

```
Title vector example : [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Content vector example : [0, 9, 1, 1, 1, 7, 0, 2, 2, 0, 0, 0, 1]
```

以上圖爬下的第三篇文章為例，可看出標題向量以上述說明，故值皆為 1，內文向量值為標題字詞在內文出現次數，若為 0 表示沒有出現在內文過。

接下來便可以套用餘弦公式來計算相似度。

程式碼如下:

```
CS = 0
for i in range(len(veclist_title)):
    # Dot and norm
    dot = sum(a*b for a, b in zip(veclist_content[i], veclist_title[i]))
    norm_a = sum(a*a for a in veclist_content[i]) ** 0.5
    norm_b = sum(b*b for b in veclist_title[i]) ** 0.5

    # Cosine similarity
    if norm_a==0 or norm_b==0:
        cos_sim=0
        Co0+=1
    else:
        cos_sim = dot / (norm_a*norm_b)
    CS+=cos_sim

avg_CS = CS/(len(veclist_title)-Co0)
print('ETtoday餘弦相符合程度:', '%.2f%%' % (avg_CS * 100))
```

```
ETtoday餘弦相符合程度: 63.16%
```

透過公式計算出 ETtoday 熱門板新聞平均相似度為 63.16%

3.4 內文情緒度分析

先將所有內文斷句，在內文中出現標點符號時斷句，接下來，將所有句子採用 SnowNLP 技術中 sentiments 函數進行情緒分析，分析的值越接近 1 表示情緒越積極正向，反之，值越接近 0 表示情緒較消極負面。

因此，以 0.5 為分界，計算大於 0.5(posi)與小於 0.5(nega)的數量，以總數除正面數量算出最終的情緒度。

程式碼如下:

```
text = str(allcontent)
stopword = [' ', '。', '、', '「', '」', '!', ';', ':']

index = [0]
for i in range(len(text)) :
    if text[i] in stopword :
        index.append(i)
lines = []
for i in range(len(index)-1) :
    start = index[i] + 1
    end = index[i+1]
    if len(text[start:end]) > 1 :
        lines.append(text[start:end])

values = []
for line in lines :
    values.append(SnowNLP(line).sentiments)

posi = 0
nega = 0
for i in values:
    if (i >= 0.5):
        posi += 1
    else:
        nega += 1

rate = posi / (posi+nega)
print('此新聞網正面用句的比例 : ','%.2f%%' % (rate * 100)) # 格式化為百分比
```

```
此新聞網正面用句的比例 :  40.14%
```

如圖可知，在該時段 ETtoday 平均文章情緒度為 40.14%

3.5 匯出至 Excel

把三個網站之每日三個時段的平均精確度、平均情緒度、及十大辭彙紀錄在 excel 檔裡，以便之後來做三個網站之總分析

程式碼:

```
wb = load_workbook(r'C:\\Users\\Yongling\\OneDrive\\桌面\\daily information.xlsx')
ws = wb.active

date = input('請輸入今天日期與時間，例：06/09 12:00：')
num = int(input('這是第幾筆資料，例：1：'))

data1 = '%.2f%%' % (avg_CS * 100)
data2 = '%.2f%%' % (rate * 100)
sortwordt = sorted(wordt.items(), key = lambda x : x[1], reverse = True)[:10]
str_w = sortwordt[0][0]
for i in range(1, len(sortwordt)) :
    str_w = str_w + sortwordt[i][0]
data3 = str_w

ws.cell(num+2, 1, date)
ws.cell(num+2, 2, data1)
ws.cell(num+2, 3, data2)
ws.cell(num+2, 4, data3)

wb.save(r'C:\\Users\\Yongling\\OneDrive\\桌面\\daily information.xlsx')
print('done')
```

```
請輸入今天日期與時間，例：06/09 12:00：06/11 22:28
這是第幾筆資料，例：1：8
done
```

Excel(截至 6/14 以前結果)：

	ETTODAY			自由時報			三立新聞網		
	精確度	情緒度	十大字彙	精確度	情緒度	十大字彙	精確度	情緒度	十大字彙
06/09 16:50	63.25%	41.94%	人曝光一年萬女死房中	63.83%	47.97%	死網手山軍台分達女報	63.77%	49.78%	字一網國了台女小大南
06/09 23:51	63.29%	41.18%	一月網光不女曝萬神了	60.75%	47.52%	大南車死公手台長不重	61.96%	48.63%	網皇了小國人來大家字
06/10 11:46	65.22%	41.53%	女一月光手大了變人曝	59.59%	43.90%	中不新出台長光公事斯	62.46%	50.38%	不國人蔣大介石這歲中
06/10 17:34	66.79%	40.96%	女人曝月歲大光公開一	60.16%	43.24%	中出不大台賴人新球有	60.17%	49.98%	蔣國不介石女字人這有
06/10 23:12	67.68%	40.14%	女快訊高一車曝光後國	62.02%	45.47%	中不大賴出一台了球高	62.10%	48.45%	蔣介石最人不後曝歲這
06/11 10:19	65.24%	40.51%	屍前一被車了高不夫妻	63.58%	42.83%	中一大國下不出天他會	61.02%	48.76%	人女大最國蔣介石不歲
06/11 14:10	64.69%	40.77%	屍一前不女曝了快訊命	65.61%	41.77%	中國一不人無了這他下	61.92%	45.91%	人大女曝國不光名都一
06/11 22:28	63.16%	40.78%	曝不女了快訊一光人大	66.03%	42.02%	中一台不點國有好重灣	64.75%	46.37%	人不大女國名最石中曝
06/12 10:05	64.97%	39.47%	不一人年台男曝被老公	56.69%	45.23%	不中了倒回台車週下一	63.19%	46.32%	不一人曝見女字看石這
06/12 14:34	63.35%	41.08%	不曝年大一了我台中女	62.04%	46.04%	不中大了台本賽市週一	61.10%	47.25%	中女國人是有台新不本
06/12 20:10	61.93%	40.79%	不一人中台了女我老快	63.10%	42.68%	中國不台大賽市防擊看	60.79%	48.06%	人一不中台有女國網了
06/13 11:00	63.52%	42.41%	女不年了一大人到星是	68.27%	40.09%	黨不賽倫市週新戰印尼	61.38%	47.75%	中爆看不了後照一網字
06/13 14:14	63.15%	42.16%	大了年女不星到一車是	66.98%	41.42%	不黨天市新黃文戲戰人	60.70%	48.37%	中了大曝爆人女後不字

3.6 結果分析

從前段程式碼中，我們得出三家新聞網的平均精確度和內文情緒度結果如下：

	ETToday	自由時報	三立新聞網
精確度	64.20%	62.68%	61.81%
情緒度	41.49%	44.21%	48.23%

從收集而來的數據統計出的平均值，可以得知近一周內三家新聞網的平均精準度大約落在百分之六十左右，其中 ETtoday 以 64.20%為三者之中較高者，推論在三家新聞網的熱門版中，ETtoday 的標題與內文的相符合程度為最高。

而近一周內的平均情緒度，三家新聞網的平均情緒度（正面用句的佔比）大約落在百分之四十至五十，整體來說偏負面新聞偏多，由此推論負面新聞可能提高新聞瀏覽量，帶來更多的收益，因此媒體刻意報導這類內容。其中三立新聞網以 48.23%為三者之中較高者，推論在三家新聞網的熱門版中，三立新聞網的內容與其他兩家新聞網相比較為積極正面。

四、研究結論

從以上研究分析，我們得出以下兩點結論：

一、若讀者經常只從新聞標題來了解新聞內容，又或者經常透過新聞標題來挑選想要閱讀新聞內容，以這三個台灣網路新聞媒體來說，我們推薦讀者選擇 ETtoday，因為標題與內文的相符程度較高，因此在讀者只閱讀新聞標題的情況下，所獲得的資訊較不會因為標題誇張的用字和用詞有錯誤的理解，也較不會在從新聞標題挑選新聞內容閱讀後，發現題文不符等情況，可以降低讀者對媒體產生的不信任感。

二、若讀者在大量閱讀較負面的新聞內容後會感到沮喪或不舒服，我們推薦讀者選擇三立新聞網，因為三立新聞網的用句較為積極正面。

五、 工作分配

王泳泠：精確度、情緒分析、report 撰寫

劉紫暄：網頁爬蟲、情緒分析、文字雲、簡報製作

廖育琳：網頁爬蟲、標題前十大字彙、文字雲、report 撰寫

六、參考資料

Python: snownlp 中文文本情感分析
https://blog.csdn.net/wangzirui32/article/details/118056830

Cosine Similarity(餘弦相似度)的計算方法及程式碼
https://clay-atlas.com/blog/2020/03/26/cosine-similarity-text-count/

[NLP][Python]透過 ckiptagger 來使用繁體中文斷詞的最佳工具 CKIP
https://clay-atlas.com/blog/2019/09/24/python-chinese-tutorial-ckiptagger/

python 處理停用詞(stopwords)
https://blog.csdn.net/miaoxiaowuseng/article/details/107343427

【Python】中文分詞並過濾停用詞
https://www.796t.com/content/1545012004.html

【wordcloud】用 Python 繪製文字雲：抓取 yahoo 新聞用 jieba + wordcloud 繪製自己文字雲看完文章 5 分鐘馬上會寫 code
https://pixnashpython.pixnet.net/blog/post/28128736-%E6%96%87%E5%AD%97%E9%9B%B2

以文字探勘技術分析台灣四大報文字風格
http://csyue.nccu.edu.tw/ch/Taiwan%20Newspapers%20(2020).pdf

Python 爬蟲實作-擷取網路新聞
https://blog.hashteacher.com/?p=1378

Python Color Constants Module
https://www.webucator.com/article/python-color-constants-module/