Final Project: Hot Topic Peeker

Student ID and name: 0710025 柯婷文

Project overview & motivation

日常生活中 google 已經成為了不可或缺的一部分，我想既然我們這麼常用，那就從我們每天 google 的東西下手好了。我發現我自己用 google 的時機其實蠻固定的:早上查天氣，聚餐時查餐廳，讀書時查學習的網站。所以我就想覺得大家使用 google 應該也會有個規律或模式，若是觀察大眾的使用模式那應該蠻有趣的。最一開始是想分析大家外送都訂什麼，就不用每次都要上網找"新竹美食"然後找到一堆廣告的食記，如果改用 google 搜尋的次數(by Google Trend)來看的話，就找越多人搜尋的越好(就算他不是最好吃的，至少會是比較多人知道的)，然後我再用 Natural Language API 去分析說搜尋這家餐廳出來的結果會是正面還是負面的，藉此判斷我要吃哪家。再來就是覺得來新竹讀書以後很少看電視，都有點跟社會脫節，所以想要利用 google trend 了解最近社會上都在討論什麼議題，再用 beautiful soup 擷取新聞內容，用 Natural Language API 分析，同時接觸正反面的資訊，才不會被媒體誤導。

Project Plan

I. 找到討論度高的話題→用 pytrends API(Google Trends' unofficial API)
   1/6 進度:pytrends 的 top chart 不能用了 改用美麗湯直接爬 google trend 的網站
   A. 使用者可以指定要搜尋的類別(如：食物/政治/電影…)
      →用 pytrends.top_charts(date, cid, geo='US', cat='')
      Cid 可以決定是要什麼主題如 basketball_players
      1/6 進度:google trend 網站沒有提供分類的熱搜排行榜
   B. 使用者可以指定要搜尋的地區(如：台北/台灣….)
      →用 pytrends.interest_by_region(resolution='CITY')等方式指定
   C. 使用者可以指定要搜尋的時間(如:每天晚上 9:00~10:00)
      →用 pytrends.get_historical_interest(kw_list, year_start=2018, month_start=1, day_start=1, hour_start=0, year_end=2018, month_end=2, day_end=1, hour_end=0, cat=0, geo='', gprop='', sleep=0)
      1/6 進度:pytrend 的這個功能也壞掉了(現在只能找 today 1-d)
II. 抓出討論度最高的十個話題
   1/6 進度:網站上每天能搜尋的資料量不一定
   →用 nlargest 函式或 list 的排序
III. 找出若是去 google 分別搜尋這十個話題 會有什麼結果
   1/6 進度:改成搜尋今天的最熱門五個話題

→用 beautiful soup 的 request.get 分別爬出標題和內文

IV. 分析出現的結果網頁，內文的評論為正面或負面以及文章的情緒

→用 beautiful soup 抓出文字 or 用 NL API 直接輸入網址(不確定可不可以)

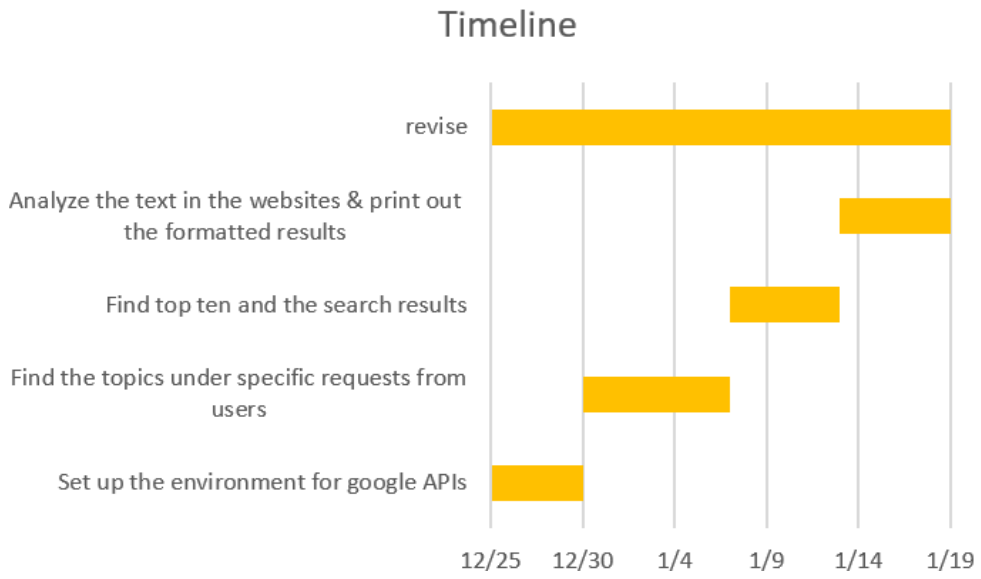
→用 Natural Language API 的 sentiment.score, sentiment.magnitude 找正面負面

V. 最後 output 出分析結果（排序第一名到第十名）

→可以用 numpy/pandas 整理成表格(或是直接用 list)，如下圖:

排名	關鍵話題 (topics)	話題熱度	情感分數 (Sentiment Score)	情感強度 (Sentiment Magnitude)	正面報導	中性報導	負面報導
1	奶茶湯圓						
2	印尼海嘯						

Timeline:



Update 1(2019/1/6)

---

1. What I have done?

找到討論度高的話題→*改用美麗湯直接爬 google trend 的網站*

註:使用者不可以指定要搜尋的類別/地區/時間

google trend 網站沒有提供分類的熱搜排行榜

pytrend 的搜尋時間功能也壞掉了(現在只能找 today 1-d)

抓出討論度最高的*今天 5 個話題跟昨天的 13 個話題*

透過 webdriver 進去網頁，然後用美麗湯找 CSS class 裡面的東西

→先用 dict 排，裡面放每個熱搜字的

link(新聞的網址，到時候就可以直接用在 natural language api)

searches(搜尋次數)

source(新聞來源)

summary(新聞標題)

title(關鍵字)

date <後面會用到>

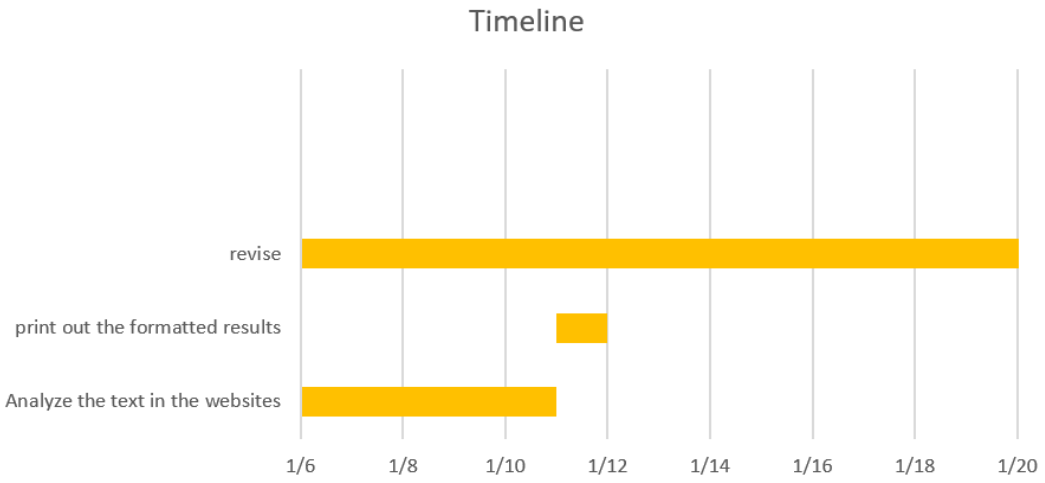
→把牠們放進同一個 list 排序

→把這個 list 轉成 panda 表格，就可以一次印出表格

2. 找出若是去 google 分別搜尋*今天的最熱門五個話題* 會有什麼結果

→用 interest by region 分析台灣五個縣市的搜尋結果差異

→印出表格



Update 2(2019/1/12)

---

1. 因為一些不可抗力的因素所以整個 project 後半部有蠻大的改變:
   a. 增加了 news api
   b. Natural language api 不能用了
2. 詳細說明
   a. 請使用者從從前面得出的 10 個結果中選出一個想要了解的關鍵字(第 0 個到第 9 個)，以及想要設定的條件(依據甚麼排序)，然後把他傳入 news api 的函式裡，就可以得到很多篇相關新聞的摘要(並有附上新聞來源跟網址)。因為這個 api 設

定是會傳回 json 格式的檔案，所以用 json 的方法把想要的資料抓出來。

b. 在抓新聞文章的時候本來想直接用爬蟲，爬每個從 pytrend 得到的網址，但是遇到很多困難，其中最大的困難就是因為每個網頁的原始碼都不太一樣，所以沒辦法寫出一個通用的爬蟲程式。後來我想說換個方法，在 yahoo(固定網站)搜尋那些關鍵字，結果竟然爬出原始碼裡面沒有的東西，完全不能理解，所以後來才改用 news api。

c. Natural Language API 的部分真的很飲恨，因為我明明都打好了，可是 google 憑證卻突然不能用了，嘗試了超多次都不行。原本是想要把上面讀到的文字內容分篇存在一個 list 裡面，然後傳入這個 API，開始進行情緒分數以及正面負面的分析。可以看出每一篇個別的分數跟總體平均相差多少，就可以觀察不同新聞媒體到底客不客觀(畢竟客觀的話應該正面負面都報導)

RUN

I. 要 install 的東西:
   A. chormedriver(https://germey.gitbooks.io/python3webspider/1.2.3-ChromeDriver%E7%9A%84%E5%AE%89%E8%A3%85.html)
   B. pip install newsapi-python
II. 流程:
   A. 在 python editor 直接 run
   B. 會彈出 google trend 視窗，等他跑完
   C. 按照指示輸入一些代碼
   D. 就可以分別得到今天和昨天的熱門搜尋字詞、某字詞的分地區搜尋熱度、依據相關程度(或熱門程度)排名的新聞其標題、新聞摘要、新聞來源、新聞網址。