GROUP 9 midterm project文字報告

- 確立動機與分析目的:起出做本專案的原因是看見有廠商提供相關的合作分析題目，這個主題對我們組員來說也都十分感興趣因此選定與智泉合作的主題「分析社會新鮮人在求職的時候在意哪些點」

- 資料集的描述:
- 資料來源:Ptt(tech job), Dcard工作版
    - job:都是徵才的文，沒有我們要的內容
    - salary:比較少人發文
    - soft job:比較少人發文
- 時間:2022/1/1 ~ 2023/1/1
- 黑名單:誠徵、徵人、徵求、求才

- 資料的分析過程:

    在最初我們先進行了資料清理，將文章進行斷詞斷句、加入停用字詞，抓出關鍵詞，建立文字雲。再來進行情緒分析，我們利用了LIWC字典法，使用CKIP,NER,POS....等方法得到情緒分數，分析每日平均情緒分數並查看該日的正面詞彙文字雲及負面詞彙文字雲。再來分析利用CKIP去跑nlp的結果(pos,ner,...)，找出每個字句的情緒分數，情緒分數的分佈，接著利用POS,NER,snowNLP後每個詞性出現的頻率彙整。再利用'研究所'當成關鍵字去找跟研究所有關的文字雲。
    最後也計算了**tf-idf**，以計算該詞彙的重要性、詞彙與詞彙之間的關聯程度，進而產生共現圖，找出字詞與字詞間的連結，以及關聯性、重要性，更可藉由連接線的粗細去判斷關聯程度。

- 視覺化的分析結果與解釋:詳見影片
- 結論:經過這些資料分析過後，發現年輕人大部分在工作版這裡的情緒分數大概都是以中性為主，再來我們發現蠻多會提到半導體、台積電這類的詞彙，那這也跟我們選的**PTT**的科技工作版及**dcard**工作版是有關係的。

    做完發文量分析之後，**2022**年發文量高峰都會是在三月到六月就是畢業前夕到下個學期開始，我們推估這個時候可能比較多求職的需求。然而，在**2022**天的四月的情緒分數比較低一點，再過一個月左右分數又回升了，推測是那個時候科技業的整體的景氣不是這麼得好。接下來我們用**CKIP**等套件去做資料分析那我們這裡一樣又是也抓到很多也是跟台積電有關係，可見台積電在台灣職場是佔有舉足輕重的地位！

    tf-idf, ngram以找出的字詞之間的關聯性，接著畫出一個共現圖，就會看到關聯性其中心點的話都是以工作為出發點，應該是因為這就是工作版。與工作有關聯的，就是像公司或是畢業或是地點**...**等這些可能就是年輕人在找工作的時候比較注重的要點。那因為台積電和人才他出現的頻率比較高，所以就特別再去看一下與他們相關性前**10**高的詞彙，可以看到與台積電最有關的詞是設廠，可能設廠設在哪邊讓大家都很關注，推測是因為他就會對於設廠處的就業人口以及帶動整體的就業景氣發展，或是房價**...**等。
- 影片連結:https://youtu.be/YLa6ni1Xsyg