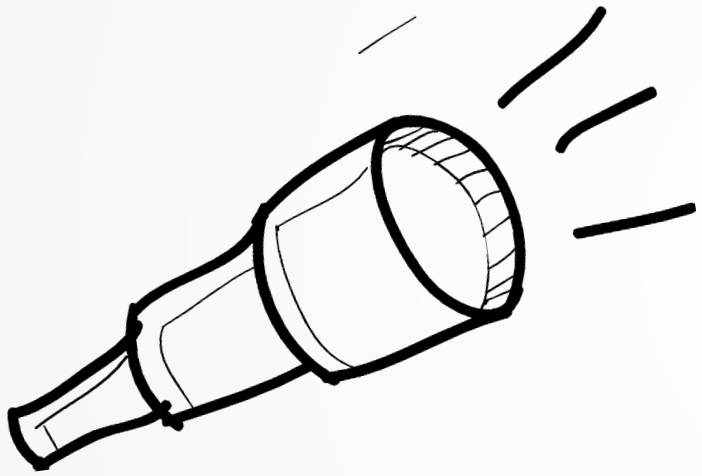
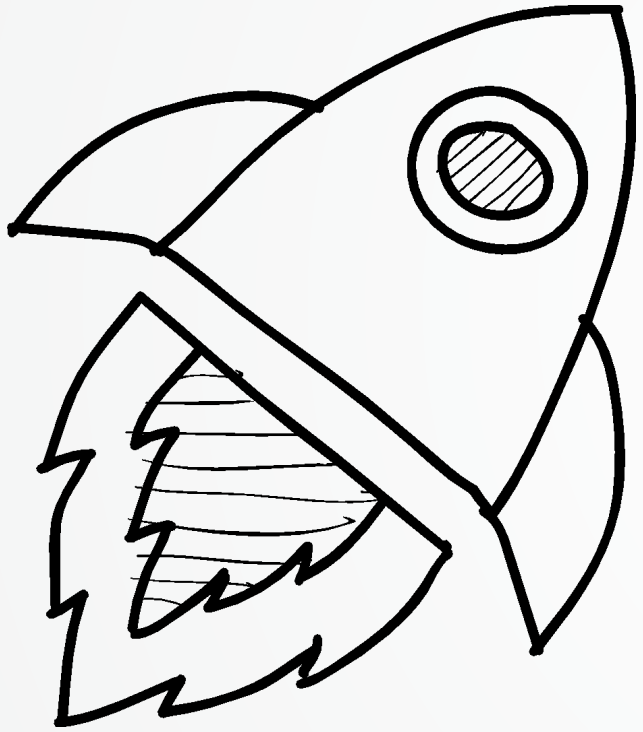
「選課沒地雷」中的情緒分析以及課程評估

陳昕淼 40621138L
陳孜旻 40621132L
顏毓廷 40640315S
黃冠瑋 40976013H

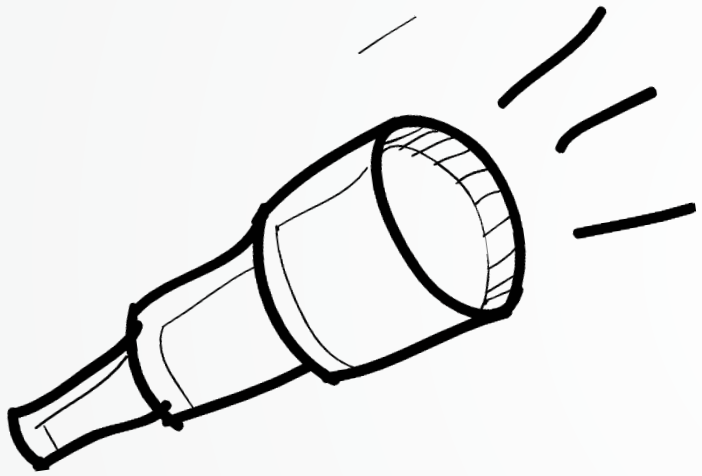


What is the problem



師大的「選課沒地雷」fb粉專中有很多同學們對於各種課程的評價，但是信息很繁雜，而且全部都是文字訊息，因此很難綜合判斷一門課程。

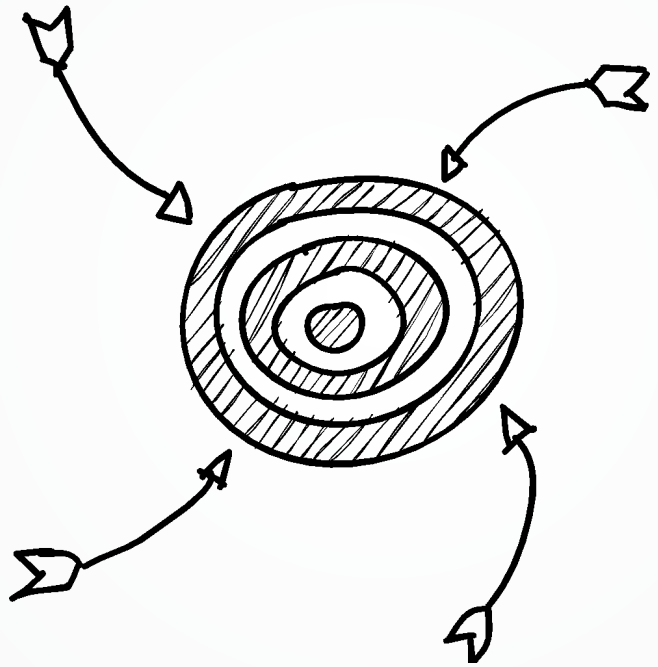
因此我們希望以「台大意見詞詞典」(NTUSD, National Taiwan University Sentiment Dictionary)為輔助，依據其中的正負面詞分類來分析「選課沒地雷」中各個課程的受歡迎程度。



Why this is important

需求，現象和問題

作為師大學生，我們每個學期選課前都會瀏覽「選課沒地雷」這個粉專來作為選課參考依據，以免選到雷課。但是我們發現裡面信息繁多，難以一下子擷取到最有用的信息；不斷地瀏覽也非常浪費時間。所以我們希望可以運用這個程式來提高大學生們的選課效率。

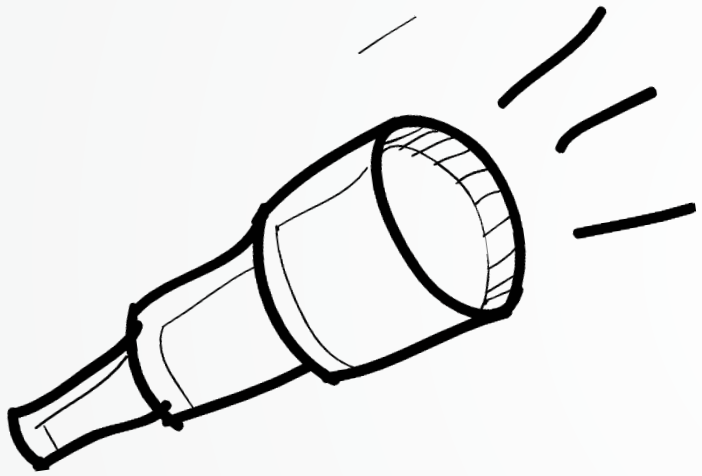


動機

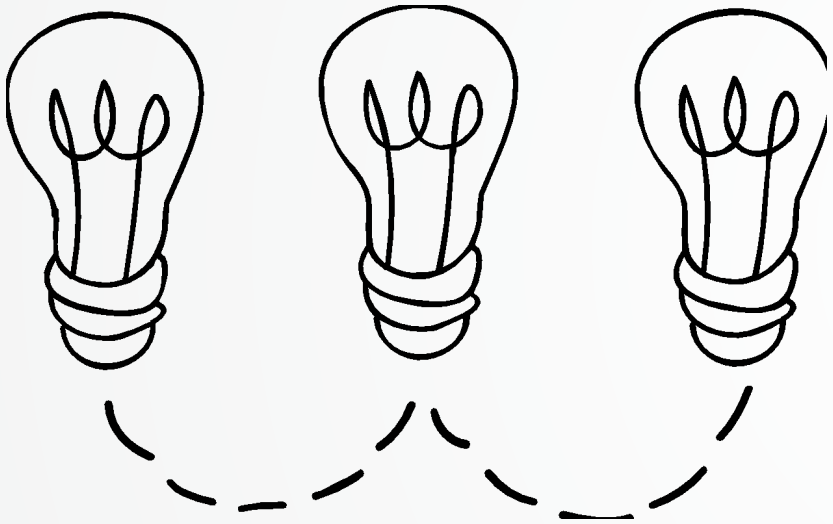
有一天，我在選課沒地雷找一節課的評價，雖然該堂課留言很多，但是正反不一，很難在看完後快速做出是否退選的決定。

為什麼值得被解決

現今臉書的情緒分析多利用按讚的分類，但此分法太過籠統。且按讚並不完全代表贊同與否。e.g. 有些人按讚只是已讀而已。



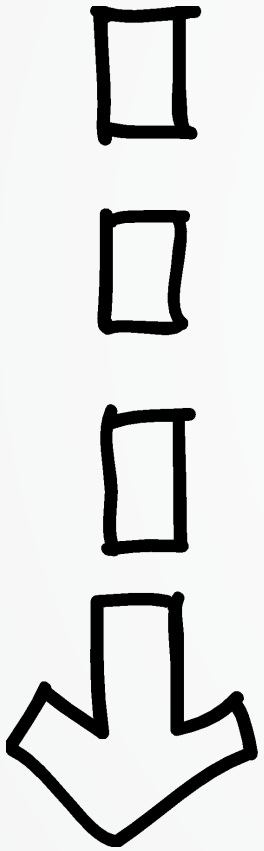
What to do and how you do it



步驟

1.收集「選課沒地雷」中某門課的文本資料
2.爬出有關課程的信息，篩選出課程下屬的留言
3.利用articut斷句，斷詞
4.逐個詞判斷是否屬於「台大意見詞詞典」中正面詞或負面詞的部分，並且具體分析特殊情況
5.分別計算正面詞個數和負面詞個數，將得分計為正面詞/負面詞，統整分數
6.得出最後的得分
7.生成文字雲

工具



Python (情緒分析程式 / 網頁爬蟲)

Articut (斷詞)

台大意見詞詞典 (正反形容詞分類)

Python (文字雲)

master ▾ final / ntusd-negative.txt

tzminch delete blank lines Latest commit cda7d7c 2 day

1 contributor

8277 lines (8275 sloc) | 78.9 KB Raw Blam

```
1  幹
2  一下子爆發
3  一下子爆發的一連串
4  一巴掌
5  一再
6  一再叮囑
7  一拳
8  一般殺人罪
9  一陣狂風
10 一陣緊張
```

負面詞詞典

正面詞詞典

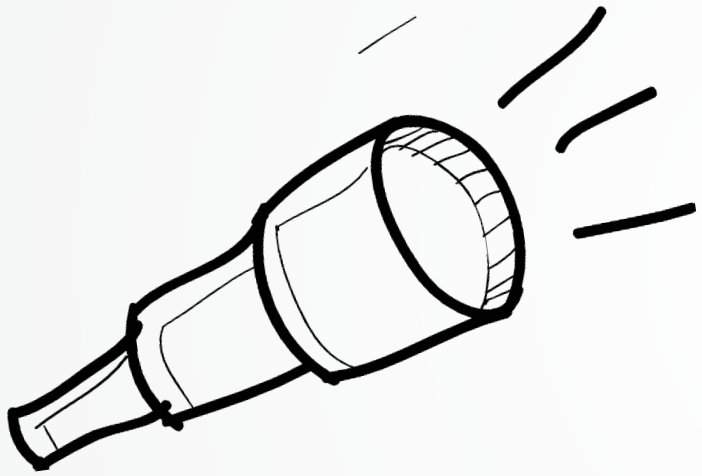
master ▾ final / ntusd-positive.txt

tzminch delete blank lines Latest commit

1 contributor

2810 lines (2810 sloc) | 25.9 KB

```
1  一帆風順
2  一帆風順的
3  一流
4  一致
5  一致的
6  了不起
7  了不起的
8  瞭解
```



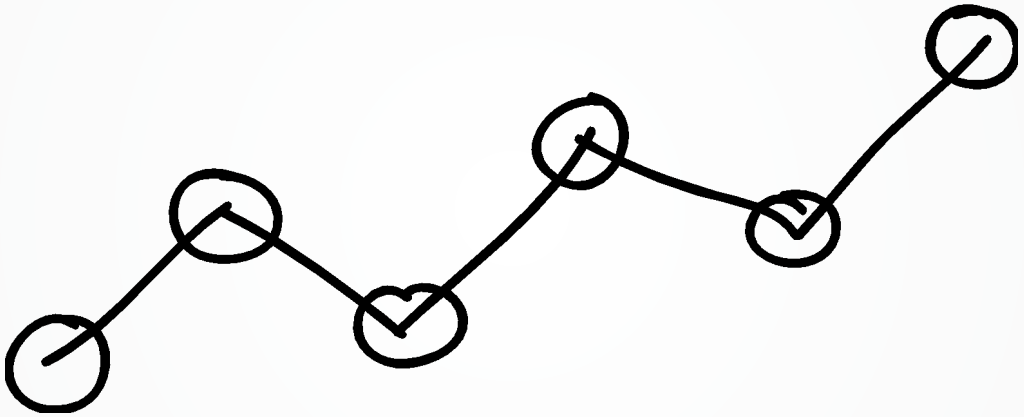
Where do the data come from

資料從哪來

「選課沒地雷」粉專

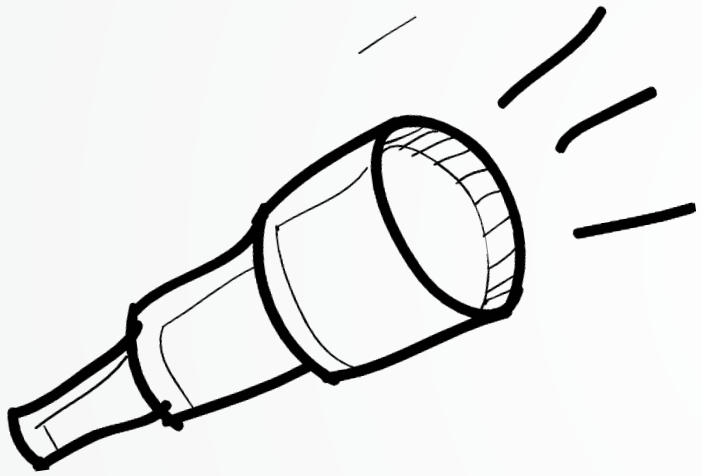
怎麼收集

用爬蟲爬



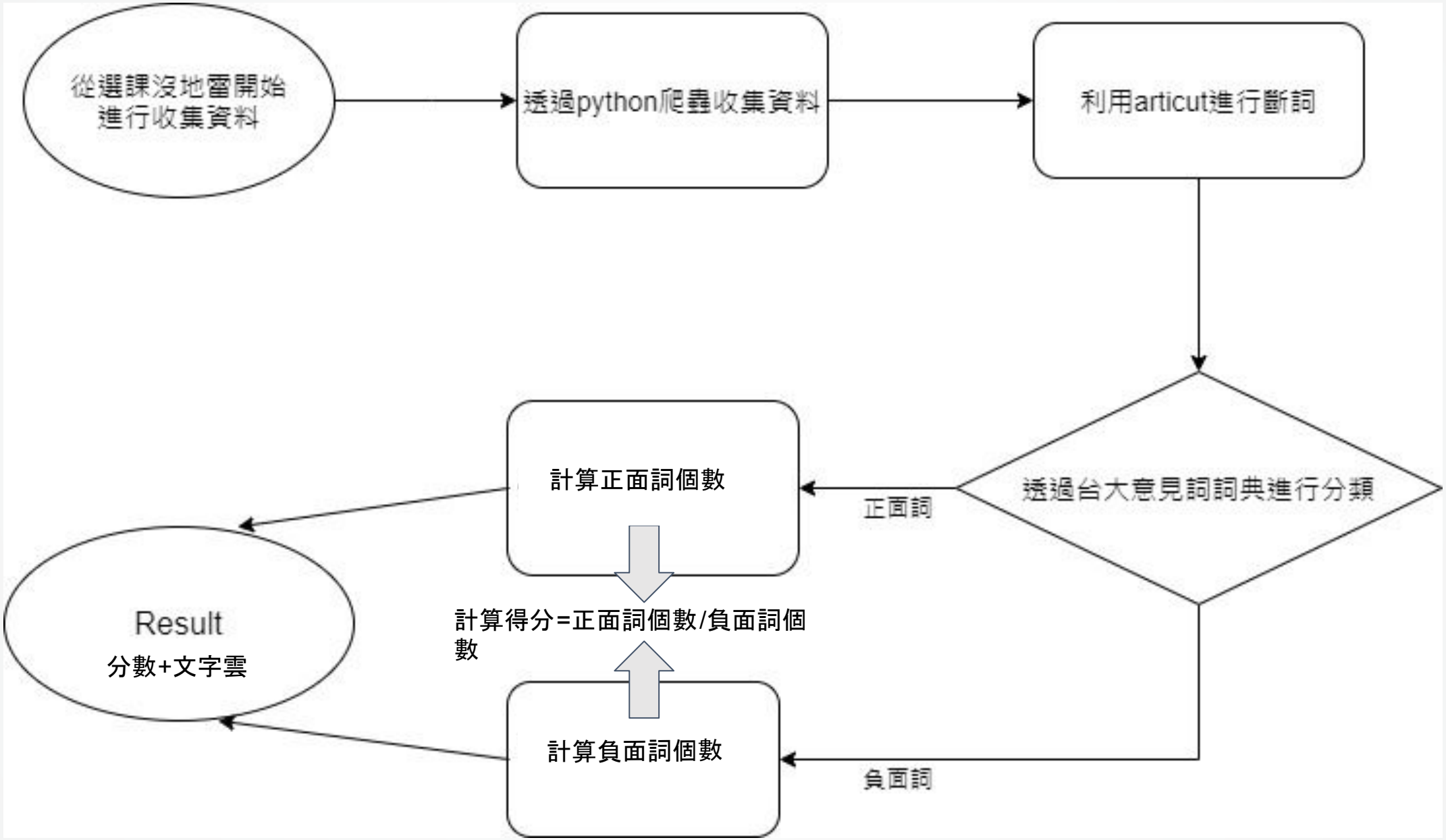
怎麼知道量夠不夠

肯定有部分冷門課程的數據不足，因此可能只對熱門課程比如通識課有更多的參考意義。綜合幾年的數據，應該會有多門課程能夠拿到足夠的樣本分析



Work flow

Github Repo:https://github.com/tzminch/final


從選課沒地雷開始
進行收集資料
透過python爬蟲收集資料
利用articut進行斷詞
透過台大意見詞詞典進行分類
正面詞
計算正面詞個數
計算得分=正面詞個數/負面詞個數
Result
分數+文字雲
計算負面詞個數
負面詞

示範文本

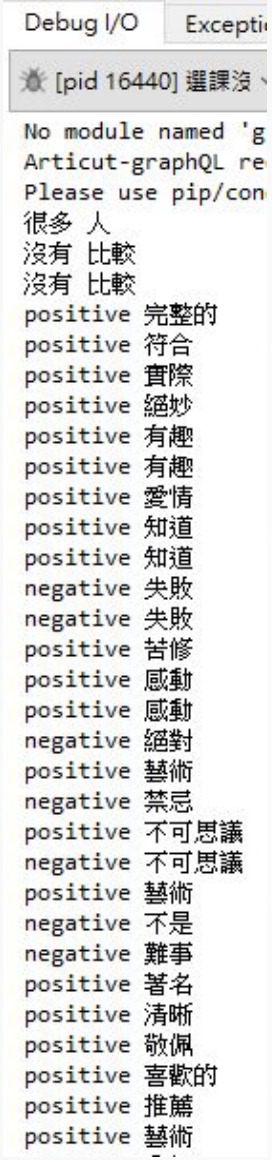
這門課很多人推，但好像沒有比較完整的介紹，來分享一下。所謂「雅文學」—是廟堂、士大夫書面的作品，與此相對的則是「俗文學」—民間口耳相傳的故事，口味符合大眾與工農兵。國高中課程多著重雅文學，以致我們對俗文學的了解甚少，實際上俗文學絕妙有趣，對文化有根深蒂固的影響。這門課會先介紹中國四大愛情傳說：春之梁祝，夏之白蛇，秋之女牛，冬之孟姜，這四則故事大家耳熟能詳，但透過老師的介紹，才會知道故事的源流與趣聞。你知道嗎？某個版本的白蛇傳中，青蛇原本是男的，追求白蛇失敗，被白蛇變成侍女，爾後白蛇被壓在雷峰塔下，青蛇苦苦修行，終於推倒雷鋒救了白蛇，白蛇大受感動，於是把青蛇變回男的，兩蛇終成眷屬。修這門課，絕對可以聽飽各種故事。介紹完四大傳說後，接著會進入各類說唱藝術：快板、評話、彈詞、相聲、琴書、雜曲.....外加猜燈謎、歇後語、民間禁忌，隨時穿插老師即興講笑話，回想起來，一學期能教這麼多內容，簡直神乎其技不可思議。沒報告、沒作業、不一定點名，期中有一個猜燈謎活動需要出席，期末則是筆試。老師會發講義，配合板書，還有播放影片跟錄音帶，需仔細聽、勤抄筆記。期末考的筆試包括填空題（四大傳說分別是？相聲的別名是？）、填充題（歇後語填寫）、聽力測驗（播上課播過的音檔，需分辨是黃梅調？彈詞？太平歌詞？）、申論題（個人對說唱藝術的感想？）拿90以上不是難事。老師本人是著名崑曲演員，講話語速極快但口齒清晰，眉飛色舞，唱作俱佳，感覺來了就現場唱一段調子，是我很敬佩喜歡的國文系教授。一門真心推薦的課，如果你對傳統藝術有興趣，來師大不可錯過。需注意，老師語速真的很快，想到什麼說什麼，板書也是類似狀況，可能會跟得有點吃力。雖然內容很廣泛，但礙於時間限制，有些會僅用影片帶過。有興趣的人會學得很高興，沒興趣大概會覺得很累...謝謝分享！之前心心念念想選這堂但都沒選到，這學期總算選到了，希望可以在課堂學到有趣的事物／／／大推孟珍老師覺得開心!我一個交換生能選到!她的課我基本上都快選完了，除了一些限修的，因為她真的很好笑，而且懂很多！她的紅樓夢也很有趣她的紅樓夢內容怎麼樣重嗎不會重喔～也是老師想到什麼講什麼紅樓夢的話要寫佳句本唷～就是看紅樓夢然後把佳句寫在自備的筆記本上，不用看完也沒關係(當初120回我才看完40回而已XD)，還會有一個上台報告，老師很free讓大家自己分組自己想要報告什麼(可以報告服飾、衣物、建築、人物的故事....)期末考筆試50格一題兩分～認真的話分數會滿好看的期末考都是她上過的內容～不用怕沒看完書林芮慶是佛心好課，不過老師上課播放的作品要認真記，是重要考點。蔡家霈

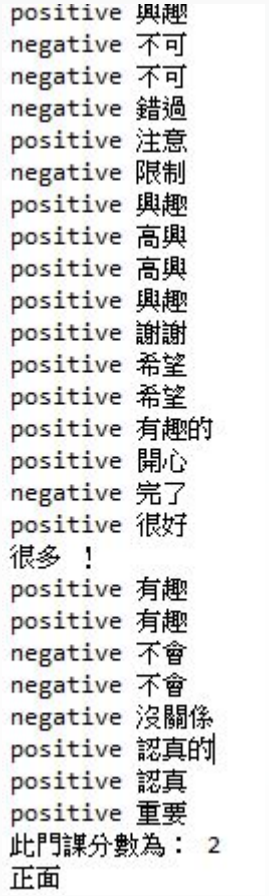
```
(top)
```

```python
#!/usr/bin/env python3
# -*- coding:utf-8 -*-

import json
from ArticutAPI import ArticutAPI
import matplotlib.pyplot as plt
from wordcloud import WordCloud

def createJson(jsonPath, inputDICT):
    with open(jsonPath, 'w', encoding='utf-8') as f:
        json.dump(inputDICT, f, indent=4, ensure_ascii=False)

def wordcloud(inputSTR):
    font = 'SourceHanSansTW-Regular.otf'
    my_wordcloud = WordCloud(font_path=font).generate(inputSTR)
    plt.imshow(my_wordcloud)
    plt.axis("off")
    plt.show()

if __name__ == "__main__":
    adjDICT = {}
    with open("ntusd-positive.txt", encoding="utf-8") as f:
        posSTR = f.read()
    posLIST = posSTR.split()

    with open("ntusd-negative.txt", encoding="utf-8") as f:
        negSTR = f.read()
    negLIST = negSTR.split()

    adjDICT["positive"] = posLIST
    adjDICT["negative"] = negLIST

    createJson('./adjectives.json', adjDICT)
```

```python
with open("民間文學與說唱藝術_禁孟珍.txt", encoding="utf-8") as f:
    inputSTR = f.read()
articut = ArticutAPI.Articut()
resultDICT = articut.parse(inputSTR, userDefinedDictFILE="adjectives.json")
#resultDICT = articut.parse(inputSTR)
#resultLIST = articut.getNounStemLIST(resultDICT)
#print(resultDICT)
resultLIST = resultDICT["result_segmentation"]
refLIST = resultLIST.split("/")

posScore = 0
negScore = 0
posSTR = ""
negSTR = ""
for i in range(len(refLIST)):
    for j in posLIST:
        if refLIST[i] == j:
            if j == "很多":
                print(refLIST[i], refLIST[i+1])
                posSTR = posSTR + refLIST[i] + refLIST[i+1] + " "
            else:
                posScore += 1
                print("positive", refLIST[i])
                posSTR = posSTR + refLIST[i] + " "
    for k in negLIST:
        if refLIST[i] == k:
            if k == "沒有":
                print(refLIST[i], refLIST[i+1])
                negSTR = negSTR + refLIST[i] + refLIST[i+1] + " "
            else:
                negScore += 1
                print("negative", refLIST[i])
                negSTR = negSTR + refLIST[i] + " "

allSTR = posSTR + negSTR
print("此門課分數為：", posScore // negScore)

if posScore // negScore > 1:
    print("正面")
else:
    print("負面")

print(posSTR)
print(negSTR)

wordcloud(posSTR)
wordcloud(negSTR)
wordcloud(allSTR)
```

結果





很多人 完整的 符合 實際 絕妙 有趣 有趣 愛情 知道 知道 苦修 感動 感動 藝術 不可思議 藝術 著名 清晰 敬佩 喜歡的 推薦 藝術 興趣 注意 興趣 高興 高興 興趣 謝謝 希望
希望 有趣的 開心 很好 很多！ 有趣 有趣 認真的 認真 重要
沒有比較 沒有比較 失敗 失敗 絕對 禁忌 不可思議 不是 難事 不可 不可 錯過 限制 完了 不會 不會 沒關係

結果 - 正面詞文字雲


知道
藝術
感動
興趣
有趣
高興
希望
很多人
完整的
不可思議
愛情
著名喜歡的
謝謝
符合
有趣的
敬佩
注意
很多
很好
絕妙
開心

結果 - 負面詞文字雲

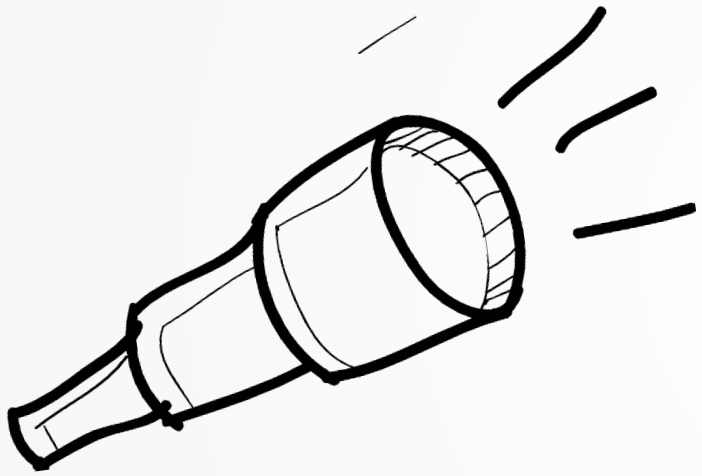
結果 - 正+負面詞文字雲


不可
希望
不可思議
敬佩
推薦
感動
認真的
雖事
重要
注意
苦修
清晰
不是
很多的
完整的
藝術
有趣
實際
不會
興趣
愛情
喜歡的
認真
高興
禁忌
著名
知道
失敗
很多人
開心
有趣的
符合
絕妙
絕對
沒關係

題外話 - 未經過斷詞的文字雲結果

問題以及改進

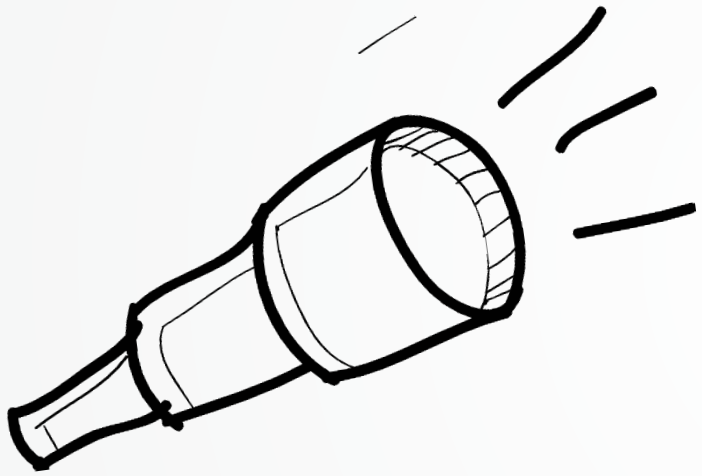
1. **中性詞如何處理？**——中性詞在「台大意見詞詞典」中同屬positive和negative兩個分類，因此在計算分數時會相互抵銷。
2. **有些詞例如「很多」、「沒有」可能並不代表正負面情緒的意思，應該如何處理？**——這些詞彙沒有很明確的正或負面情緒傾向，所以將它們改定義成中性詞。我們最後的改進是：當遇到類似詞彙時我們將其設定為不計分，然後抓取出該詞後面的一個詞出來，將其顯示在最後的結果中讓大家自行參考判斷。
3. **有些語助詞可能屬於負面詞但是不一定用來表達負面意思，應該如何處理？**
4. **最後的得分是否會被資料數量所影響？**——在之前的程式中，在資料量不同的情況下兩門課不能依據得分高低來判斷是否更受歡迎，但是可以通過得分的正負來判斷該門課程是受到好評還是差評。所以當時我們主要判斷的依據不是得分的高低而是正負。但是後來改進成百分比的形式，受資料數量的影響大大減少了，分數也具有可參考意義。



How it could be used in the future

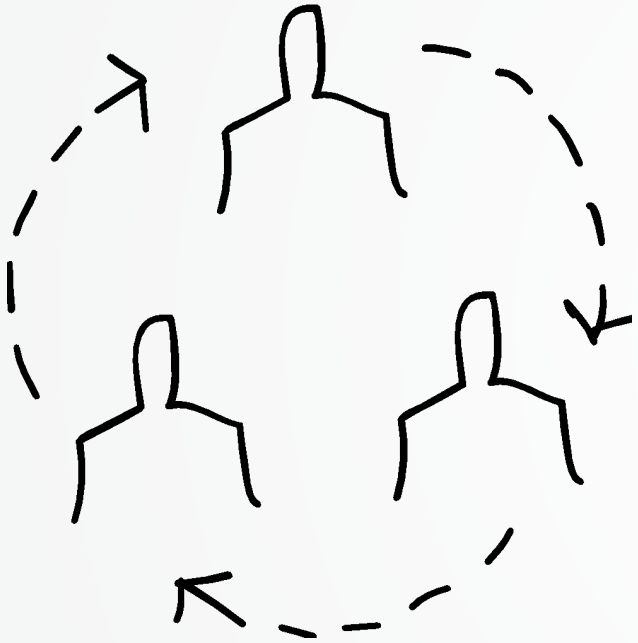
應用

1. 用於不同FB社群的情緒分析
2. 用於分析不同產品的用戶評價



Who did what in your team

分工

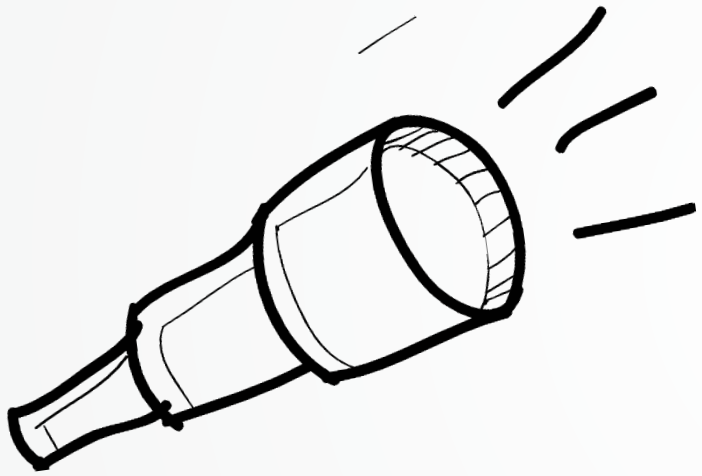


陳昕淼：簡報及文案、報告、小組討論

陳孜旻：程式設計(正反面詞分析+文字雲)、簡報、報告

顏毓廷：小組討論、報告

黃冠瑋：



Q&A

THANK YOU