**人工智慧期末報告**

**U1033125**

**周庭安**

**側翼粉專判斷器**

**主題說明:**

**現在的台灣充斥著各個政黨的水軍側翼炒作等等亂象，為 了令大家好判斷粉絲專業的政治傾向，故寫此側翼粉專判斷 器。**

**內容說明:**

**首先找尋幾個已經被確定抱出來的FB側翼粉專藍綠政黨共四位，我找的分別為國民黨:「藍色力量」與「反蔡英文粉絲團」，民進黨:「只是堵藍」與「謝立聖插畫」，經由爬蟲爬出他 們粉絲專業內的文章，並標註1 or 2 的標籤，經由羅吉斯回歸判斷器去訓練模型，**

**最後讓使用者將自己有疑慮的粉專內文貼上判斷，即可判斷政治傾向。**

**使用套件: csv jieba pands sklearn joblib beautifulsoup driver webdriver time codecs os**

**程式碼-FB爬蟲:**

**from bs4 import BeautifulSoup**

**import driver**

**from selenium import webdriver**

**from webdriver\_manager.chrome import ChromeDriverManager**

**import time**

**import csv**

**import codecs**

**driver = webdriver.Chrome()**

**#driver.get("https://www.facebook.com/%E5%8F%AA%E6%98%AF%E5%A0%B5%E8%97%8D-245311342852891/")**

**#driver.get("https://www.facebook.com/XieLiSheng/")**

**#driver.get("https://www.facebook.com/ROC.Blue.Power/?ref=page\_internal")**

**driver.get("https://zh-tw.facebook.com/people/%E5%8F%8D%E8%94%A1%E8%8B%B1%E6%96%87%E7%B2%89%E7%B5%B2%E5%9C%98/100064877342886/")**

**#滑**

**for x in range(300):**

**driver.execute\_script("window.scrollTo(0,document.body.scrollHeight)")**

**#print("滑")**

**time.sleep(2)**

**html = driver.page\_source**

**soup = BeautifulSoup(html, 'html.parser')**

**# 定位文章標題**

**titles = soup.find\_all("div", class\_="x78zum5 xdt5ytf xz62fqu x16ldp7u")**

**#csv寫入**

**csv\_writer = csv.writer(codecs.open('BLUE-2.csv', 'w', 'utf-8'))**

**for title in titles:**

**posts = title.find\_all("div", dir="auto")**

**if len(posts):**

**for post in posts:**

**csv\_writer.writerow([0,post.text])**

**這段程式碼是在爬取粉絲專業的內文，比較重要的點為 FB為動態網頁，內容不會一次跑出來，必須要有滑鼠滾輪滑一下的動作，**

**for x in range(300): 的迴圈便是在做這件事，透過了javascript 的辦法來解決，投過迴圈多次滑動，便可爬取大量內容，最後在html找我要的東西，輸出成csv檔案。**

**程式碼-模型訓練:**

**import csv**

**import jieba**

**import pandas as pd**

**from sklearn.feature\_extraction.text import CountVectorizer**

**from sklearn.linear\_model import LogisticRegression**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn import metrics**

**import joblib**

**data = []**

**data\_label = []**

**#讀入 分詞 轉換格式**

**with open('bgdata.csv', 'r',encoding='UTF-8') as csv\_file:**

**csv\_reader = csv.reader(csv\_file)**

**for r in csv\_reader:**

**data.append([' '.join(jieba.cut(r[1]))])**

**data\_label.append(r[0])**

**df = pd.DataFrame(data, columns=['article'])**

**df['label'] = data\_label**

**#切分測資**

**data\_train,data\_test,data\_label\_train,data\_label\_test = train\_test\_split(df['article'], df['label'], test\_size=0.2)**

**#轉換器**

**vectorizer = CountVectorizer()**

**data\_train = vectorizer.fit\_transform(data\_train)**

**data\_test = vectorizer.transform(data\_test)**

**joblib.dump(vectorizer, 'vectorizer.pkl')**

**#邏輯**

**BGmodel = LogisticRegression()**

**BGmodel.fit(data\_train, data\_label\_train)**

**#測**

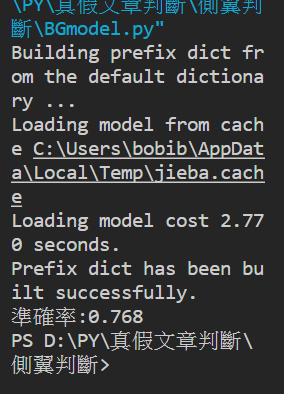
**pred = BGmodel.predict(data\_test)**

**s = metrics.accuracy\_score(data\_label\_test, pred)**

**print("準確率:%0.3f" % s)**

**joblib.dump(BGmodel, 'BGmodel.pkl')**

**將爬取資料製成的資料檔進行斷句分詞轉換等等操作，變成邏輯思回歸器可以接受的資料，並與作業相同分出一些資料來做測試，最後透過羅吉斯回歸器訓練模型。**

****

**程式碼-使用者輸入:**

**import jieba**

**from sklearn.model\_selection import train\_test\_split**

**import joblib**

**import os**

**model = joblib.load('BGmodel.pkl')**

**vectorizer = joblib.load('vectorizer.pkl')**

**while 1 :**

**text = input('請輸入要判斷的粉絲專業內文:')**

**if text!='@@':**

**text\_cut = [' '.join(jieba.cut(text))]**

**p = vectorizer.transform(text\_cut)**

**pp = model.predict(p)**

**if pp[0]=='1':**

**print('這偏綠')**

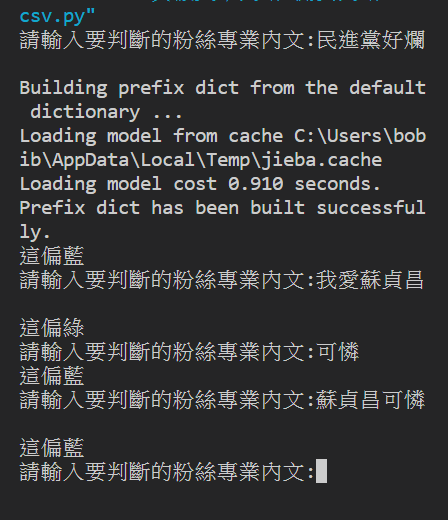
**elif pp[0]=='0':**

**print('這偏藍')**

**elif text=='@@':**

**os.\_exit()**

**將建立好的模型匯入後，讓使用者輸入想判斷的粉絲專業內容，依樣對此內容做格式的轉換等，最後輸出答案。**

****

**專題的特色及製作時碰到的困難:**

**在製作本次專題時一開始想得很美好，但在製作時遇到了幾個問題:許多的粉絲專業其實內容只會有一張圖片搭配一些簡短的文字，圖片還是多個不同但同為側翼的粉絲專業一起發出(圖一)(圖二)，我能爬取的內容很少也容易沒頭沒尾的，造成了我的模型準確率不高，或許之後的努力方向是可以加如圖片文字的辨識，提高準確率。另外老師給的問題 其一是指輸出「可憐」二字，判斷為偏藍的，要解決這個問題，因為是讓使用者自行輸入，但要求的是輸入粉專的內文，除非是有圖片不然不太有可能只有可連二字，其二是由於我示範輸入了「我愛蘇貞昌」答案為偏綠，老師想判斷是否只要有蘇貞昌就為偏綠，於是叫我輸入「蘇貞昌可憐」來判斷，答案為偏藍，代表我的模型訓練的還算可以，但我也認為這是很有可能發生的，所以未來或許可以導入褒貶慈辭典與政治詞彙辭典來對斷句判斷等等做更好的分析來解決這問題。**

****

**圖一 圖二**

**這整學期的學習心得及建議:**

**人工智慧程式設計是一門非常有趣的課程，我們學習了一們新的程式語言，與許多有關模型訓練、爬蟲、深度學習的知識。學到了如何使用 Python實作各種算法，並學習實作了許多厲害有趣的套件，來達成目標。**

**透過學習人工智慧程式設計，我深刻地了解到人工智慧對於未來社會發展的強大引響力，例如最近很紅的OPEN AI製作的Chat GPT 與我電腦上有裝的AI繪畫stable-diffusion-webui軟體等等，都是代表著科技進步的結晶。**

**這門課非常有挑戰性，但我很享受學習的過程，也很高興我能夠學到這些知識。希望這些知識將會在未來中發揮重要作用，而我也會繼續鑽研罩門領域。**