與該領域的專業人士匹配職業建議問題

-資科final project

統計系111  
H24071011  
 曹舒瑜

統計系111  
H24071029  
 常思為

統計系111  
H24076029  
 劉米婷

1. 背景與動機

我們在kaggle中尋找適合作為final project 分析預測資料的dataset，最後找到data-science-for-good-careervillage這筆資料，我們認為此資料的數據量十分充足，有許多不同文字數字和類別的特徵，會利用許多處理文字的套件，也能使我們充分利用在資料科學導論這門課上所學的資料分析模型，不管是類別資料或是多變量的回歸，再進一步延伸至機器深度學習的觀念，藉此找出對此資料分析預測最有利的結果，來呈現在這一學期中我們在資料科學導論這門課所學到的相關概念與應用，尚且在現今競爭激烈的社會中，這些將面臨升學問題以及出社會找工作的年輕人來說，內心總是有許多的疑問，這時候就需要有一群富有經驗的人員來替他們解答問題，因此我們選擇此題目來做分析預測，希望找出最能有效解決這些年輕人疑難雜症之辦法。

1. 問題描述

CareerVillage.org是一個提供那些缺乏經驗的年輕人尋找職業或是未來方向的一個網路平台，我們的研究目的便是替那群年輕人找出最適合為他們解答的專業人士，再將問題匹配給他們，而用來評斷問題是否適當匹配的標準，來自當年輕人收到來自專業人士的答覆後，對回答所給出的評論分數，每一筆由年輕人提出的問題都會標上一個特定的標籤，因此也能適當運用各個標籤去找出最適配的專業人士來回答問題，因為必須使年輕人的問題匹配給最適合回答問題的專業人士，在此網路平台中，所有的專業人士都需要經過一定的審核，才可進入此平台作為解決年輕人提問的人員。

1. 潛在應用與價值

我們透過資料科學來從資料中分析出各筆資料所著重的特徵，透過不同模型的交叉分析，找出最適合這筆資料的分析預測模式，使用大數據將這種常態會不斷出現的職業生涯問題找出最適當的回答，解決多數年輕人所碰到的問題，我們使用簡單的模型，卻透過分析資料，探索更多的問題，找出最適合得模型比例後，就能使資料處理和預測自動化，達到資料分析預測更有效率的真正價值。

1. 資料描述

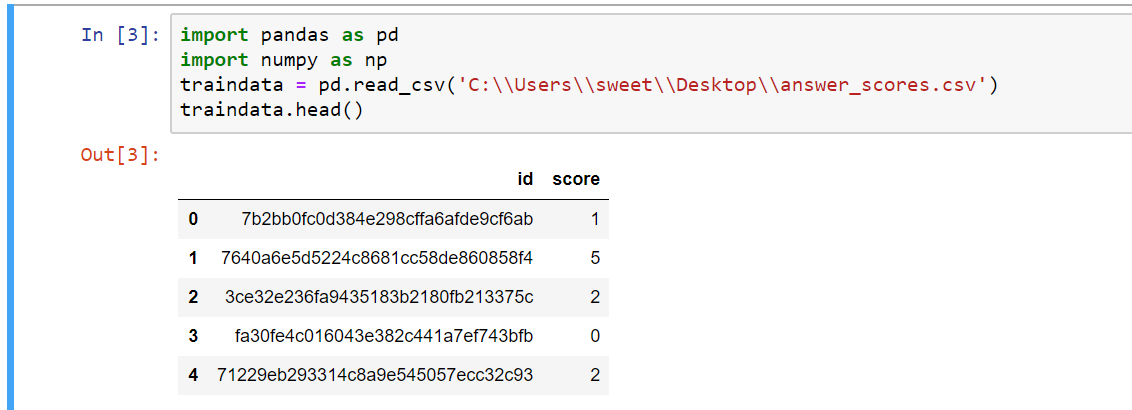
answers.csv

這筆資料內容為專業人士對年輕人對職業與生涯的問題敘述所做出的回答和相關時間id等等



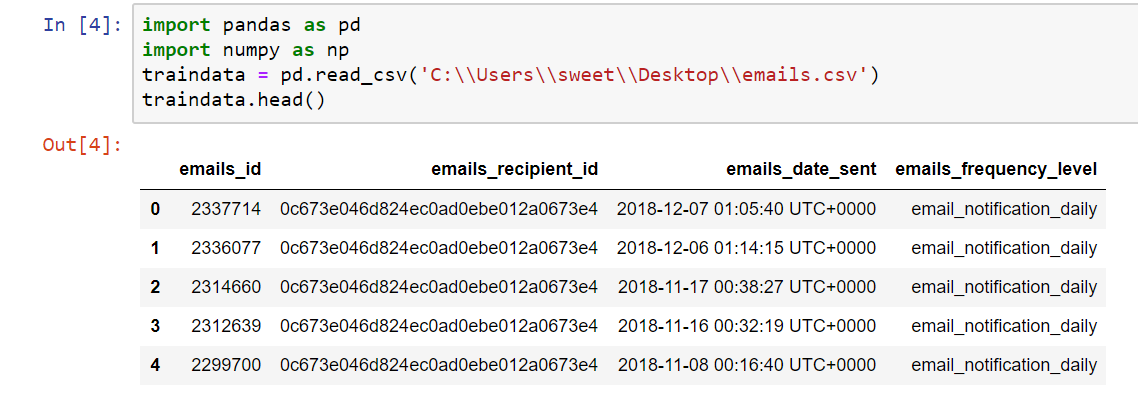
answer\_scores.csv

此筆資料代表年輕人對所接收到的答覆所給出的辦別水準分數



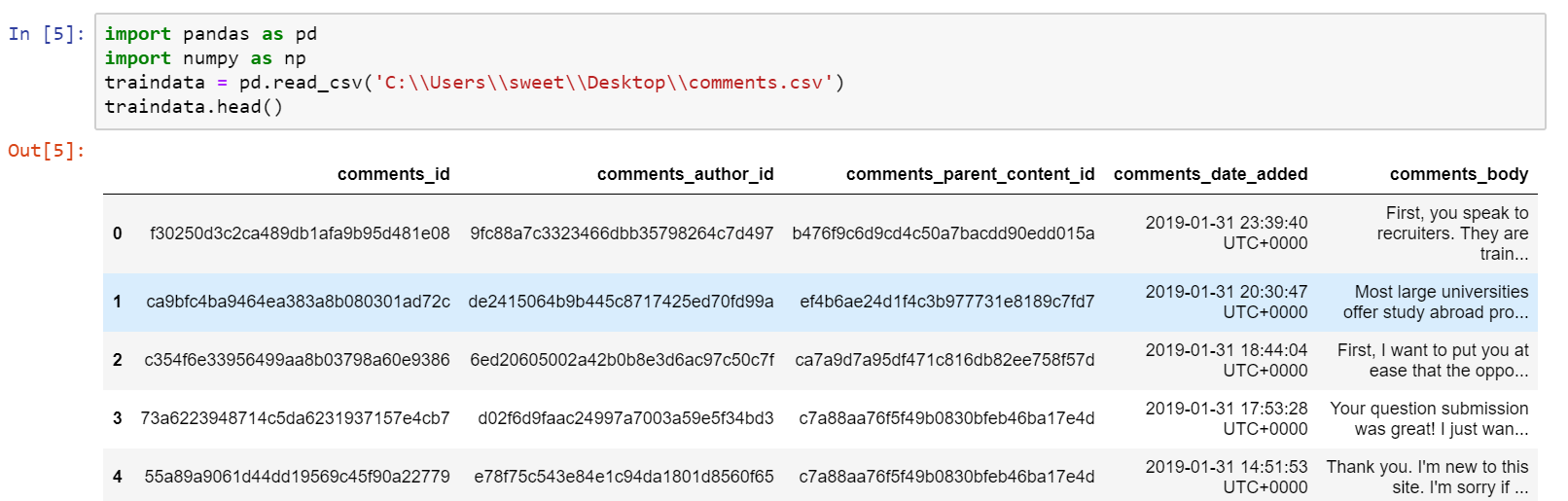
emails.csv

此資料為提問者的emails與提問頻率類別



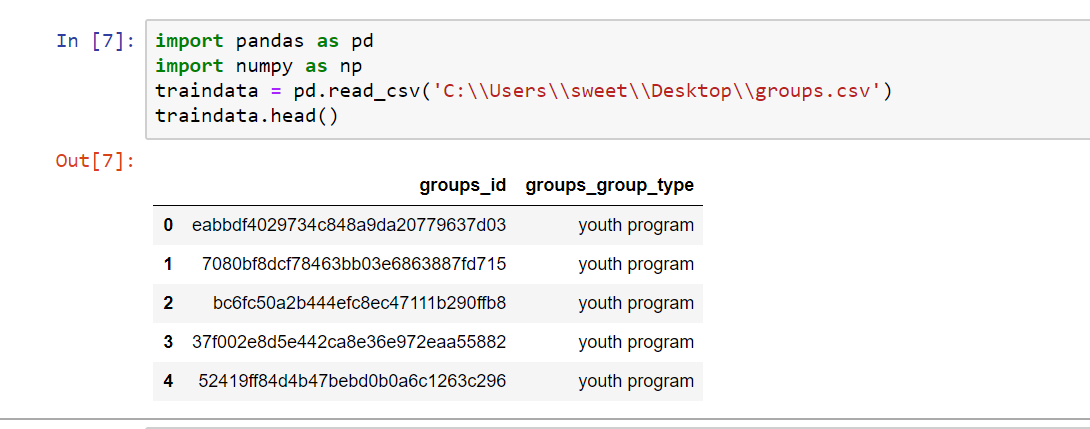
comments.csv

此資料為來自提問者收到問題答覆後所做出的文字評論



groups.csv

此資料為提問年輕人所屬哪類型的相關背景

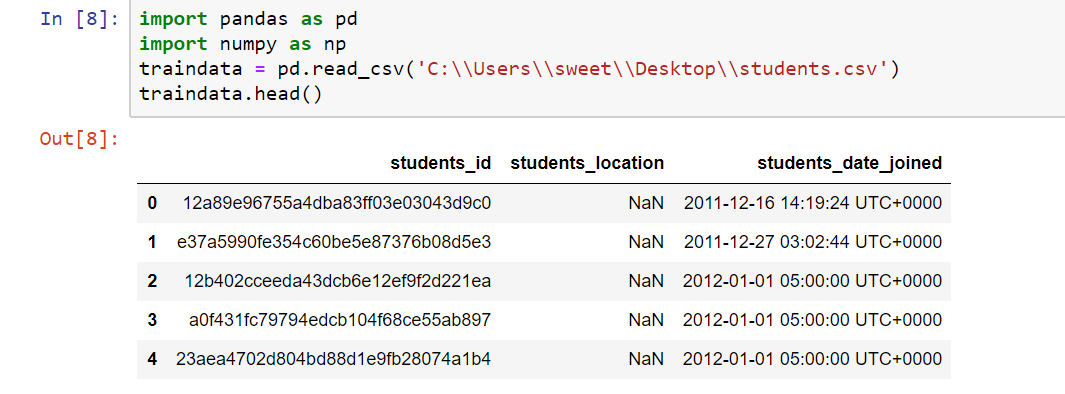


group\_memberships.csv

school\_memberships.csv

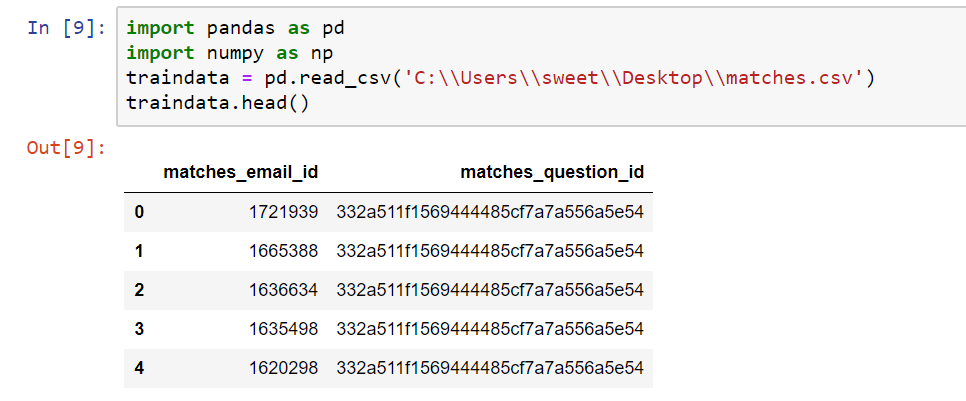
students.csv

此資料包含提問者的id與居住地區等等相關



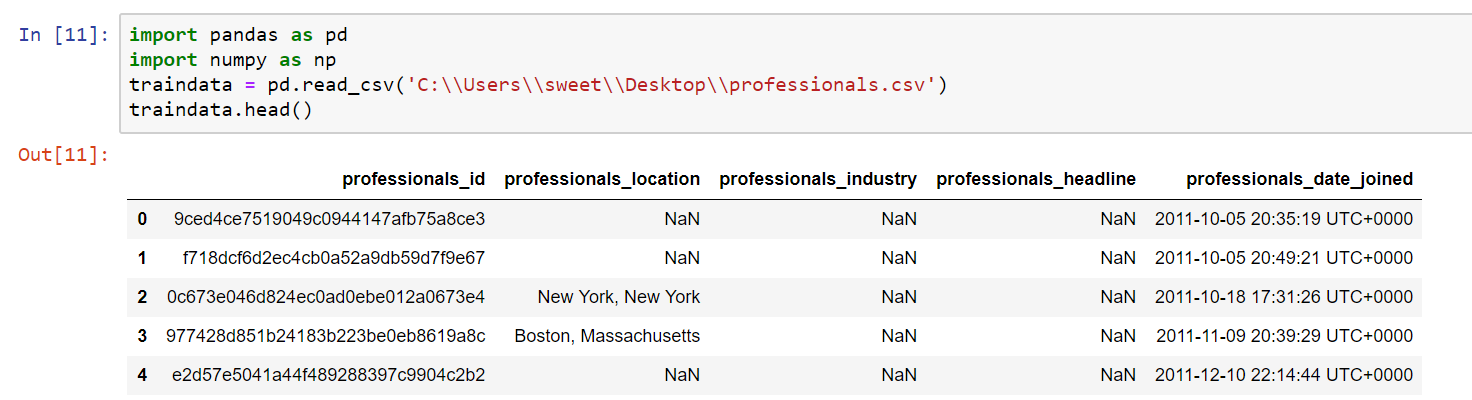
matches.csv

此資料會提供電子郵件中包含哪些問題。如果電子郵件只包含一個問題，則該電子郵件的ID將僅顯示在此處一次。如果電子郵件包含10個問題，則該電子郵件的ID將在此處顯示10次。



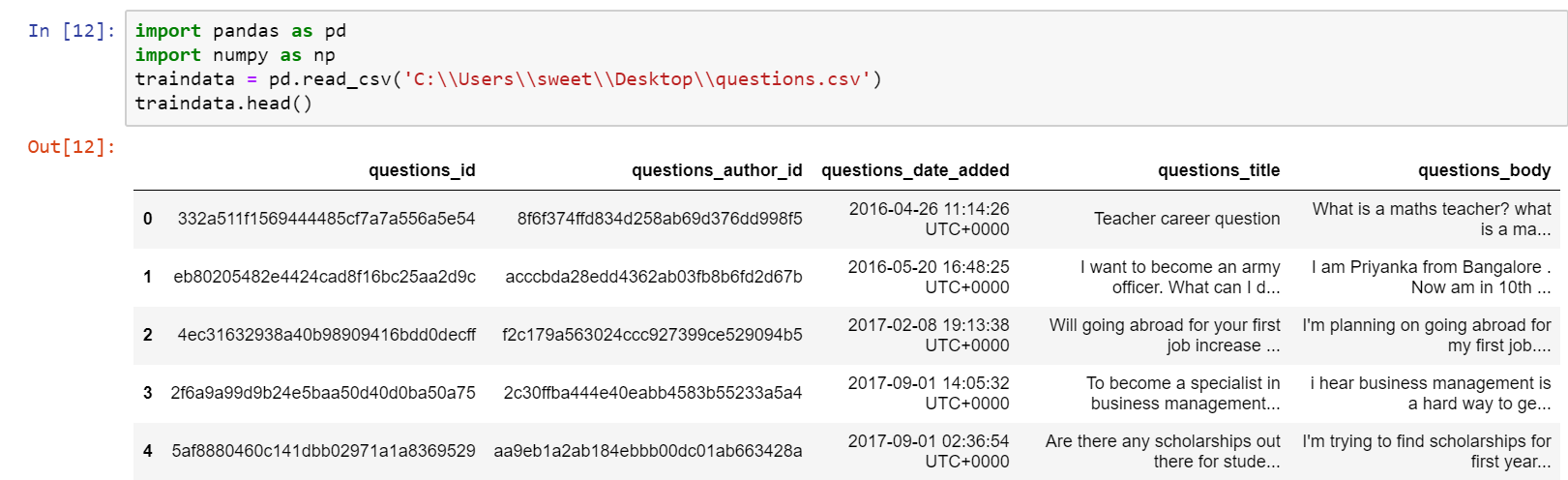
professionals.csv

此資料代表回答問題的專業人士所屬哪類別的專業領域與其基本居住等等資訊



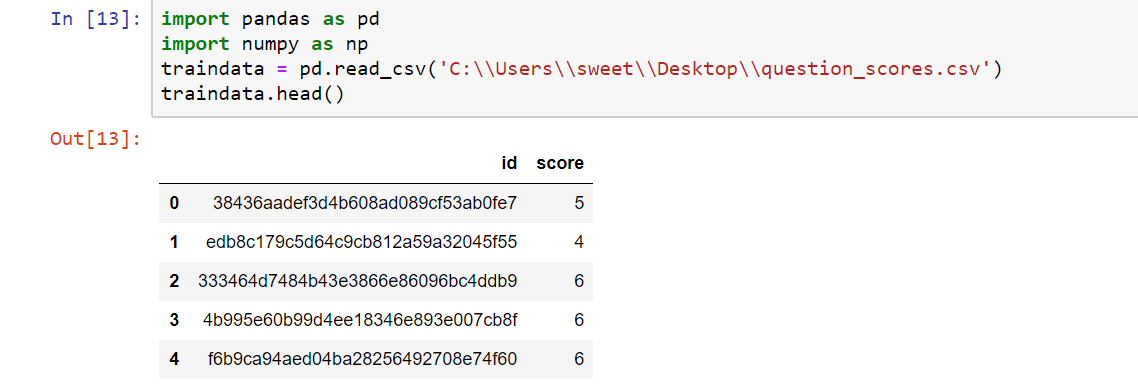
questions.csv

這筆資料內容為年輕人對職業與生涯的問題敘述和相關時間id等等



question\_scores.csv

此筆資料代表對於提問者所提出之問題的程度辦別分數



tag\_questions.csv

tag\_users.csv

tags.csv

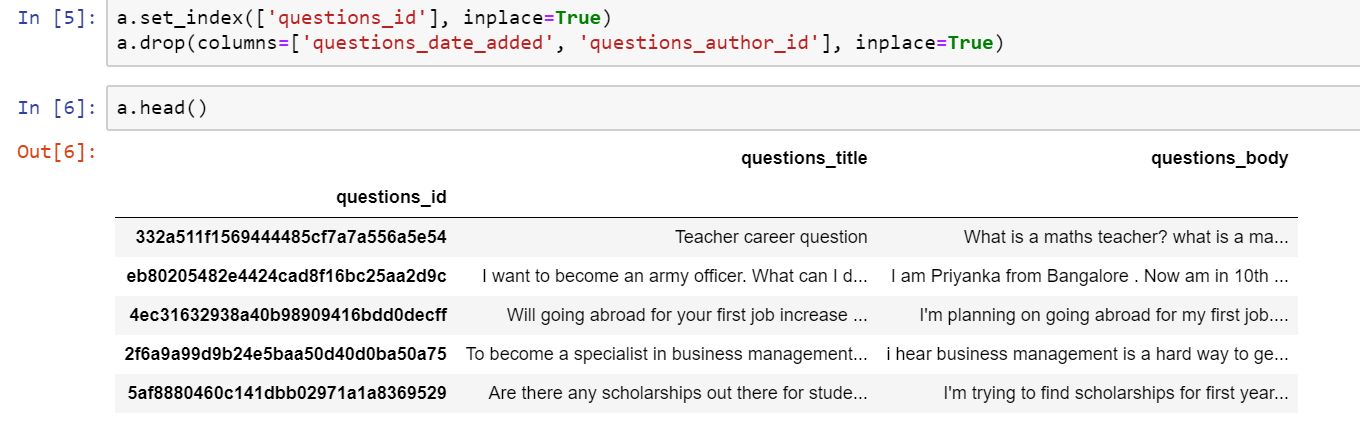
此資料為以利於將年輕人的問題做出更好的分析結果而產生的問題標籤分類



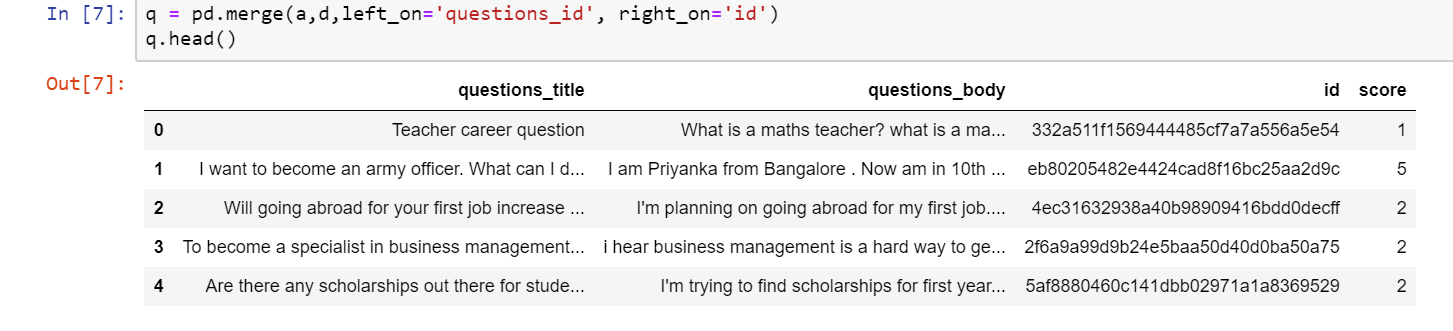
1. 具體作法

資料前處理:

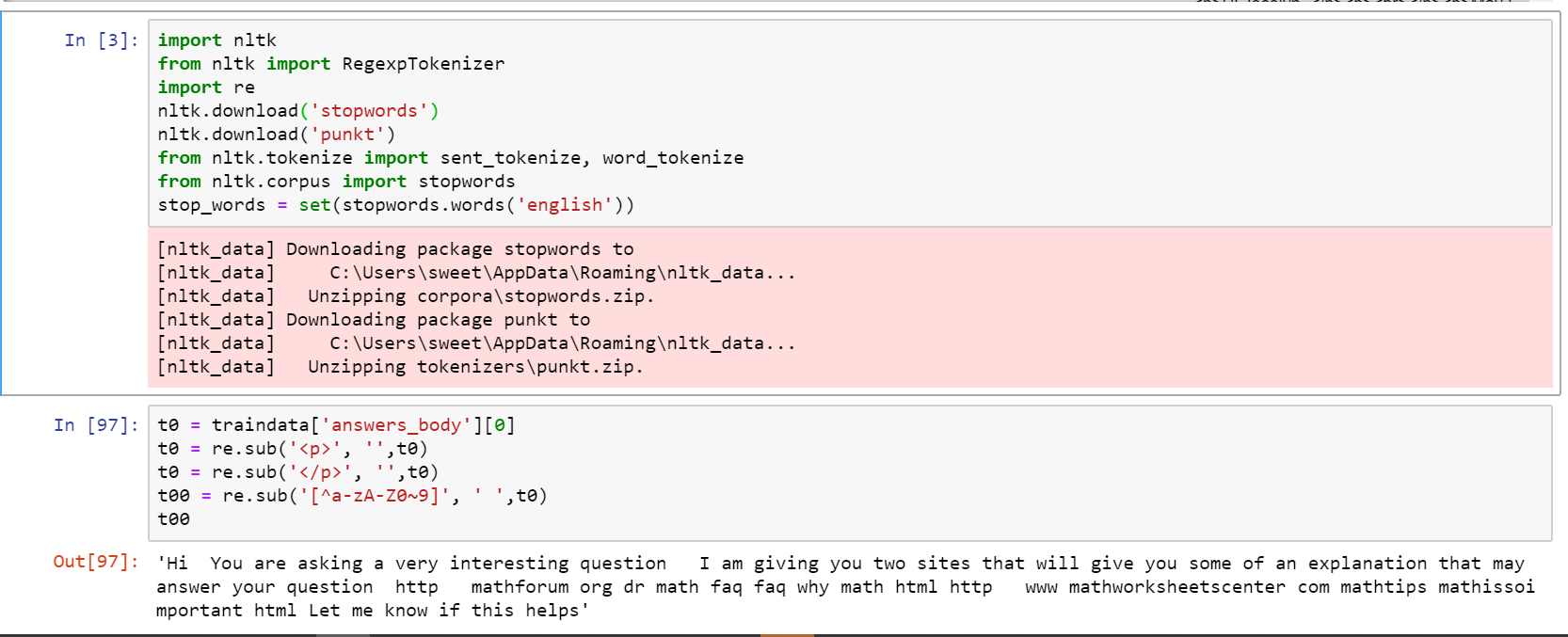
我們先去除一些我們認為不會對資料預測結果產生很大影響的特徵

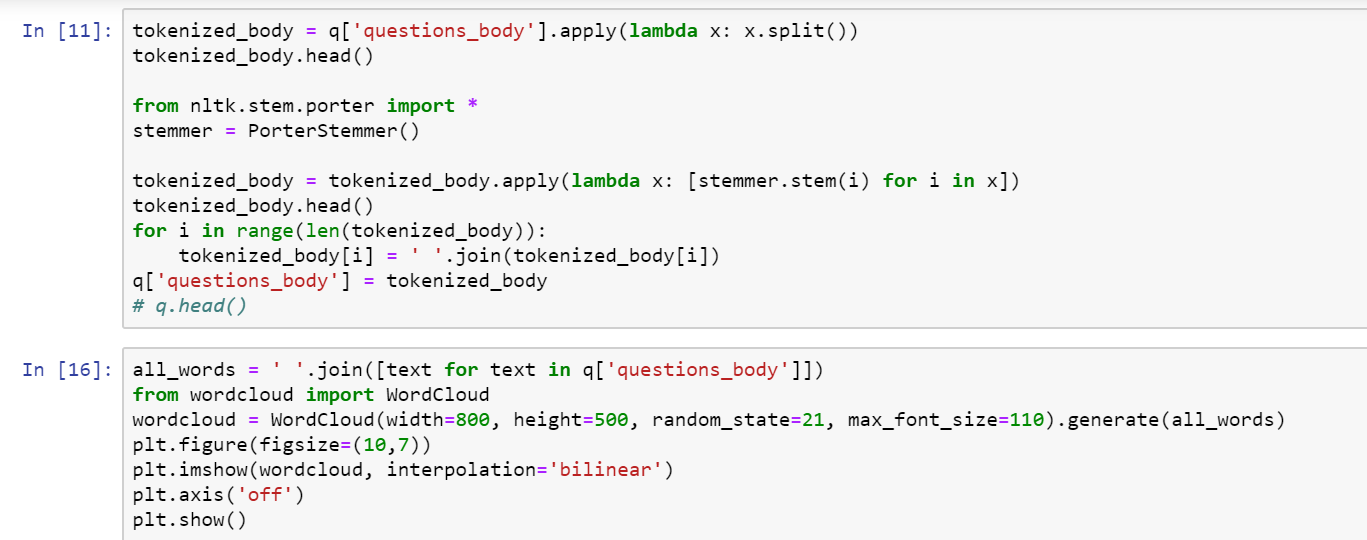


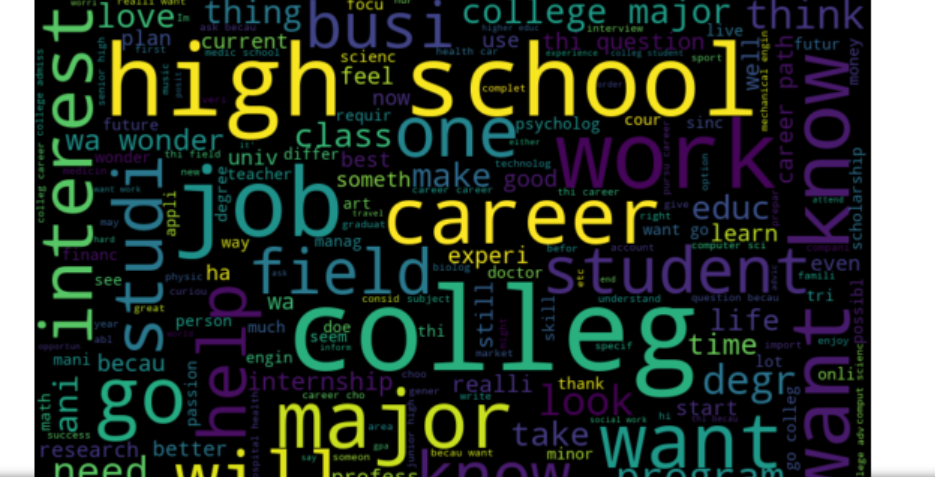
接著用merge把各個含有相同id的資料放入一個表格，因為同一個id可能會有不只一筆的資料，同一個年輕人可能提出多個不同的問題



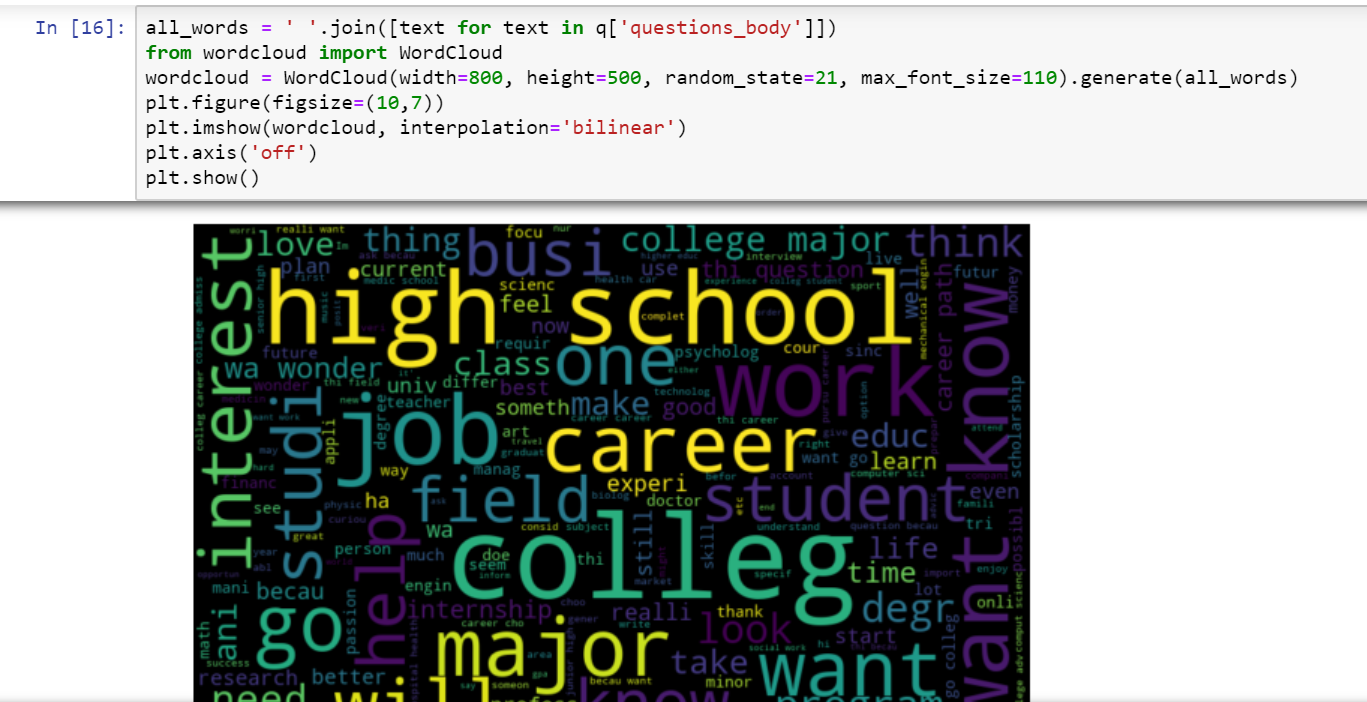
再用nltk處理文字 例如:stopwords,tokenize等等







接著分析各個資料彼此之間的關係，在分析資料的過程中，我們也找出了提問著者所提出問題中前十大關鍵標籤，例如:大學升學問題、修課問題、職業走向問題、醫藥相關問題與商業問題等等



在處理文字當中，我們也運用tfidf將文字分析的更完整，計算各個字詞與問題之間的相關度，去加強分析文字後預測的效果



接著我們將整個資料切割成測試資料與訓練資料，在放入我們所使用的模型:

LogisticRegression

LinearRegression

PolynomialRegression(多項式回歸)

經過調參數與交叉分析之後得出最後的預測結果

1. 模型評估與實驗結果分析

預估使用模型:

RandomForest

Xgboast

Catboast

KNeighborClassifier

MLPClassifier

SVM

LigjtGbm

LogisticRegression

Neural Network

實際運用模型:

LogisticRegression

LinearRegression

PolynomialRegression(多項式回歸)

由於這筆資料的預測結果為0~10的分數，不同於類別資料所分析出的結果，我們選擇利用MSE以及MAE來評估此次資料預測結果的準確性。

1. 結論與心得

統計111 常思為

這次的final project題目就讓我們頭很痛，想說給自己一個挑戰，但是因為各個資料之間複雜的關係讓我們在分析各個資料之間的關係和資料前處理的部分花了很大的心力，尤其之前接觸到的大多是一對一的資料，但是這次遇到的狀況是一對多的資料！我覺得這次的final project讓我發現到原來一筆資料的狀況是可以那麼複雜的，彼此間交錯的關係，非常需要耐心細心地去觀察。

統計111 曹舒瑜

在剛開始尋找適合final project的題目時，我們先選擇了一個預測心臟病得病機率的題目，但其資料與cp1的分系類型十分相似，我們這組認為都已經學到學期快要結束了，卻還在作和學期初一樣的事情，有點失去final project的意義，因此我們重新搜尋kaggle的資料庫，最後終於找到了這個預測問題匹配度的資料，此筆資料的數據量足夠，也有多個特徵值以利我們做更準確的分析以及預測，但在一開始，我們就花了一些時間在讀懂每個資料檔中各項特徵所代表的意義，還有評斷此預測資料準確度的方法等等，先搞清楚之後，讚開始處理資料中的文字，經過斷詞處理與將不必要的文字去除後才開始做真正的分析，切割訓練資料與測試資料，再嘗試套入不同的模型，這次的困難點是我們在資料前處理上花了許多時間將各筆分散的資料串成同一筆再找出不通特徵之間的關聯等，再加上我們的時間所剩不多，因此並沒有得出一個很理想的結果，希望有機會還是能花一下空閒的時間把資料分析與資料科學的各種模型再好好了解與應用。

統計111 劉米婷

這次的Final Project挑戰不是單純的文字處理，而是文字以及其他資訊的關係，因此必須在各個分散的資料檔中找到關聯。剛開始想要嘗試的模型有很多，但後來發現模型的使用也有很多的條件，所以其實真的有跑出來的模型只有幾個。比較可惜的是我們還沒有徹底地將所有的資料帶入考慮，因為有些不知道應該如何尋找關係的數據我們就先捨棄，有看到其他人的做法是用Neural Network及其他深度學習的模型，如果有機會想要多花一點時間研究一下這些模型要怎麼使用。這學期學了很多的分類方法也學會帶很多模型，但我覺得自己並沒有真的了解所使用的模型所跑出來的結果有甚麼意義，我覺得這是我未來可以再思考然後努力的方向。