R12725049徐尚淵 作業三

1. 執行環境：VS code
2. 程式語言：Python (版本3.11.4)
3. 執行方式：

一張含有 文字, 字型, 螢幕擷取畫面, 電子藍 的圖片

自動產生的描述直接使用VS code GUI Run code

1. 作業處理邏輯說明
   1. 讀取需使用的package

一張含有 文字, 螢幕擷取畫面, 字型, 行 的圖片

自動產生的描述

* 1. 用Array與List的方式紀錄Training data的資訊

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

* 1. Tokenize Function (同作業一)

一張含有 文字, 螢幕擷取畫面, 軟體 的圖片

自動產生的描述

* 1. 讀取Stopwords並設計前處理function (同作業一)

一張含有 文字, 螢幕擷取畫面, 軟體 的圖片

自動產生的描述

* 1. get\_unique function用於將重複字過濾掉

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

* 1. Vocabulary function

首先先將訓練文件id以及字彙存成一個numpy array，再利用While迴圈計算字彙出現的次數，最後以 numpy array的資料型態return出去。

一張含有 文字, 螢幕擷取畫面, 軟體 的圖片

自動產生的描述

* 1. 計算文件數以及將字詞整合的Function

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

* 1. 將文件中Token抓出來的Function

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

* 1. TrainMultinomialNB

利用老師講義的虛擬碼做修改，V以及N就利用前面所寫過的Function Vocabulary以及CountsDoc來實作取得，分別是一個array以及數字，V\_Selected則是透過設計計算的Chi\_Square的Function所得出的500個重要字，接著對每個Class做迴圈，其中又包含兩個針對500個重要字的迴圈計算個字詞屬於某class\_id文件中的條件機率。

一張含有 文字, 螢幕擷取畫面, 軟體, 作業系統 的圖片

自動產生的描述

* 1. ApplyMultinomialNB

利用老師講義的虛擬碼做修改，利用 TokensFromDoc 函式獲取文檔的詞彙列表，將每個class的先驗機率的對數添加到score中，並計算條件機率，最後return分數最高的類別，因array參數從0開始因此最後要加1

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

* 1. Feature Selection Function >> 用Chi Square來實作

cal\_single\_chisquare：小的 chi 矩陣，該矩陣包含一個類別的 present 和 absent 數，以及移除掉目前類別的總和，緊接著計算各格的期望值以及chi-square，累加過後產生最終的chi-square值

cal\_square\_test：對於每個training data set 的class和Doc，獲取其詞彙列表，並對每個每個詞彙，計算其 chi-square 分數，最後return前 500 個 chi-square 分數最高的詞彙作為重要特徵

一張含有 文字, 螢幕擷取畫面 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面 的圖片

自動產生的描述

* 1. 主程式部分：讀取檔案、訓練、測試以及輸出

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述