IR HW2 Report

B05705025 資管四 陳漢威

1. 執行環境:

VSCode, terminal

2. 程式語言

Python3

3. 執行方式

程式會讀入 training.txt 與裡面包含的 training data，在同一資料夾內路徑為”IRTM”的資料夾，並開啟裡面所有的檔案作為計算的依據，程式最終會產出 csv 檔案，包含 id 與 value 兩項

該次使用的套件有(編譯前須先 pip install):

- nltk (for Stemming and stopwords)
- re (for filtering the symbol)
- numpy (for vector operation)
- collections (for terms count)
- pattern (for lemmatization)

4. 作業處理邏輯說明:

讀入路徑中所有文字檔的同時會進行前處理與切詞、詞性還原等動作，並利用 Counter() 去計算不同 category 的 document 中 terms 集合的加總。

之後利用 chi square method 作 feature selection，選擇最高的 500 個 feature 作為計算機率的 terms。

計算完之後再依據 terms 在 category 中出現的次數去計算 conditional probability，並存成 dictionary。

最後，再依據我們計算出來的機率丟入 testing data 作比較，再用 argmax 去 predict 該 documents 在哪個 category 的機率最大，選擇最大的作為我們的 output。

最後的 output.csv 大致長相:

Id	Value
17	2
18	2
20	2
21	2
22	2
23	2