112學年度 數值方法期末報告

期末報告名稱: Email Classifier

姓名：許評詔

指導老師：游濟華

<設計理念>

隨著數位通訊的普及，電子郵件已成為日常溝通的重要工具。然而，每當信箱用久了，難免會收到許多垃圾郵件。

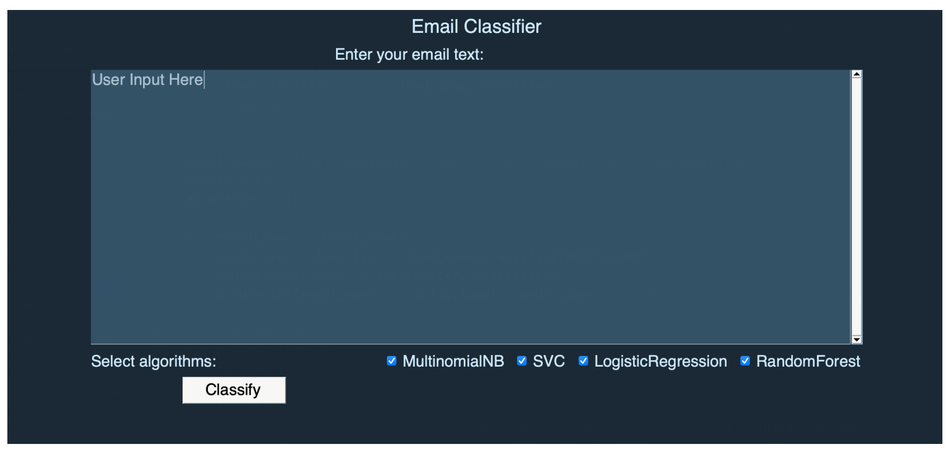
旨在開發一款智能垃圾郵件分類器，能夠自動識別並過濾垃圾郵件，提高使用者的郵件使用體驗和效率。

<設計流程>

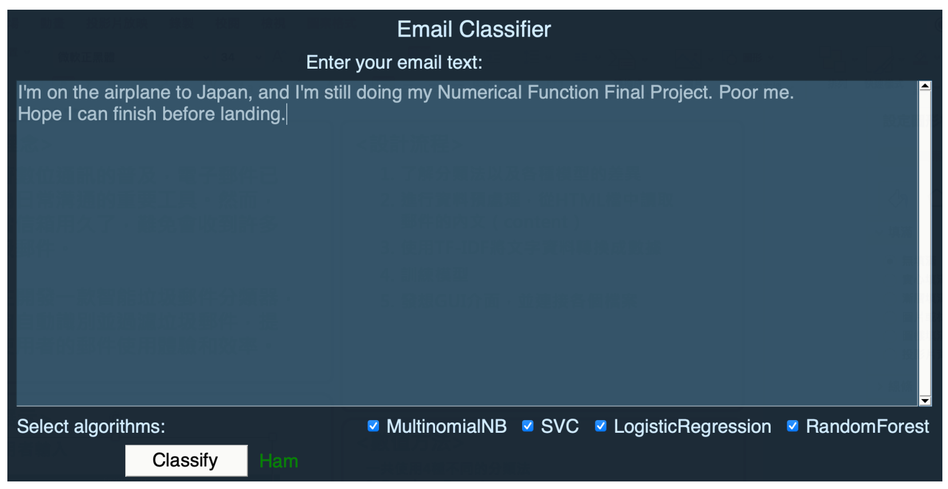
1. 了解分類法以及各種模型的差異
2. 進行資料預處理，從HTML檔中讀取郵件的內文（content）
3. 使用TF-IDF將文字資料轉換成數據
4. 訓練模型
5. 發想GUI介面，並連接各個檔案

<成果展示>

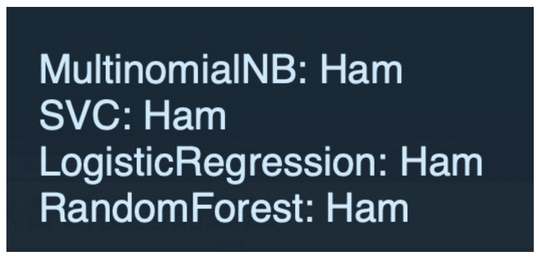
1. 使用者輸入



2. 選擇模型後進行分類



3. 查看各個模型的結果



<未來展望>

- 連結Gmail API，提升實用性
- 優化介面，使用其他的GUI模組
- 加入深度學習的模型

<數值方法>

1. Multinomial Naive Bayes (多項式朴素貝葉斯)

多項式朴素貝葉斯是一種基於貝葉斯定理的概率分類器，特別適用於文本分類問題。它假設特徵（即詞語）之間是獨立的，特徵出現的次數服從多項分布。

•計算條件概率：對於每個類別，計算每個詞語在該類別中出現的條件概率。

•預測：給定一封新郵件，計算該郵件屬於每個類別的後驗概率，並選擇概率最大的類別作為預測結果。

2. Support Vector Classifier (SVC，支持向量分類器)

•尋找最優超平面：SVC 通過尋找一個能夠最大化類別間距的超平面來分隔數據點。這個超平面使得兩類數據點之間的間隔最大化。

•支持向量：支持向量是離超平面最近的數據點，它們對確定超平面起關鍵作用。

•核函數：這裡使用線性核作處理

3. Logistic Regression (邏輯回歸)

•線性模型：邏輯回歸使用線性函數將輸入特徵映射到一個概率值，表示某個樣本屬於特定類別的概率。

•邏輯函數：通過邏輯函數（Sigmoid 函數）將線性輸出轉換為介於 0 和 1 之間的概率值。

•閾值決策：根據預設的閾值（通常為 0.5），將概率值轉換為二分類的預測結果。例如，大於 0.5 的概率值預測為垃圾郵件，反之為正常郵件。

4. Random Forest (隨機森林)

•構建多個決策樹：從訓練數據中有放回地隨機抽樣多個子集，並為每個子集訓練一棵決策樹。

•節點分裂：在每個決策樹的節點處，隨機選擇部分特徵並根據這些特徵分裂節點，以減少數據的方差。

•集成預測：通過多數法將多個決策樹的預測結果結合起來，作為最終的分類結果。這種方法能降低單個決策樹過擬合(overfitting)，並提高分類的準確性。

<相關連結 & 參考資料>

Youtube Demo: https://youtu.be/vuCVJEdP3kA

GitHub: https://github.com/KendellHsu/Spam-Email-Classifier/tree/main

Train Dataset : https://www.kaggle.com/datasets/beatoa/spamassassin-public-corpus/code