程式語言識別與分類

自然語言處理與文件探勘

小組期末專題說明文件

組員

資工三 109590004 呂育瑋

資工三 109590037 歐銘耘

**環境設置**

我們使用Visual studio code來執行main.ipynb，環境建置依vscode官方執行ipynb的推薦方式進行設定，並額外安裝以下package：

pip install scikit-learn

pip install pandas

pip install openpyxl

**與報告不同修改處**

原本是以讀取由kaggle下載的snippet.db，經由SQL語法取得資料後直接進行資料處理；後改由從snippet.db讀取資料並寫入對應的文字檔，再讀取文字檔進行資料處理，目的為避免執行程式需讀取db檔案耗時、避免需下載snippet.db(解壓縮後檔案為60GB)，改以多份txt檔案後容量較小，並修改對應資料讀取程式，其餘程式碼與流程不變。

一張含有 文字, 螢幕擷取畫面, 軟體, 字型 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 軟體, 多媒體軟體 的圖片

自動產生的描述