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。

資料提取(無須執行，已產生檔案)

```
import sqlite3

conn = sqlite3.connect('../snippets/snippets.db') #僅第一次執行須下載檔案
cursor = conn.cursor()
code_type_list = ('Bash', 'C', 'C++', 'Go', 'Java', 'JavaScript', 'PowerShell', 'Python', 'Ruby', 'Rust')

for code_type in code_type_list:
    cursor.execute(f"select snippet from snippets where language = '{code_type}' limit 10000;")
    rows = cursor.fetchall()
    with open(f'../snippets/{code_type}.txt', 'w', encoding='utf-8') as f:
        for row in rows:
            f.write(row[0])
    print(f"{code_type}.txt create successfully.")
```

[2] ✓ 0.4s

資料清洗

```
import random

x_train, y_train, x_test, y_test = [], [], [], []

for code_type in code_type_list:
    with open(f"../snippets/{code_type}.txt", 'r', encoding="utf-8") as f:
        lines = f.readlines()
        rows = [''.join(lines[i:i+5]) for i in range(0, len(lines), 5)]
        random.shuffle(rows)
        train_size = int(len(rows) * 0.8)
        train_data = rows[:train_size]
        test_data = rows[train_size:]
        label = code_type
        for data in train_data:
            clean_data = remove_annotation(data, label)
            if clean_data != "":
                x_train.append(clean_data)
                y_train.append(label)

        for data in test_data:
            clean_data = remove_annotation(data, label)
            if clean_data != "":
                x_test.append(clean_data)
                y_test.append(label)
```

[3] ✓ 4.3s