**Natural Language Processing and Text Mining: HW#3**

**小組成員和負責工作：**

**109590041 范遠皓 撰寫程式和測試**

**109590043 柯瑞霖 撰寫程式和報告**

**環境**

使用的語言：Python

所需套件：

gensim==4.3.1

joblib==1.2.0

numpy==1.24.3

pandas==2.0.1

python-dateutil==2.8.2

pytz==2023.3

scikit-learn==1.2.2

scipy==1.10.1

six==1.16.0

sklearn==0.0.post5

smart-open==6.3.0

threadpoolctl==3.1.0

tzdata==2023.3

安裝辦法：

Zip檔內有requirements.txt

終端機輸入以下指令安裝

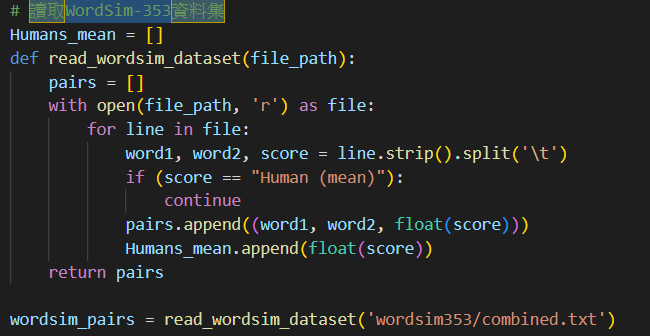
pip3 install -r requirements.txt

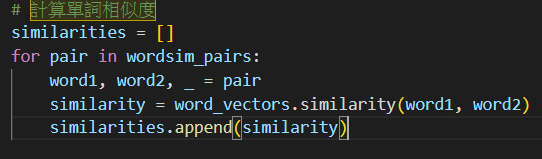
載入預訓練的詞向量模型，使用 gensim word2vec model

[下載連結](https://www.kaggle.com/datasets/leadbest/googlenewsvectorsnegative300)

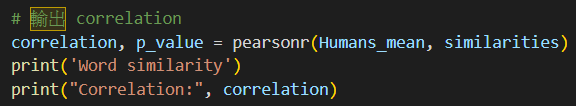


讀取WordSim-353資料集

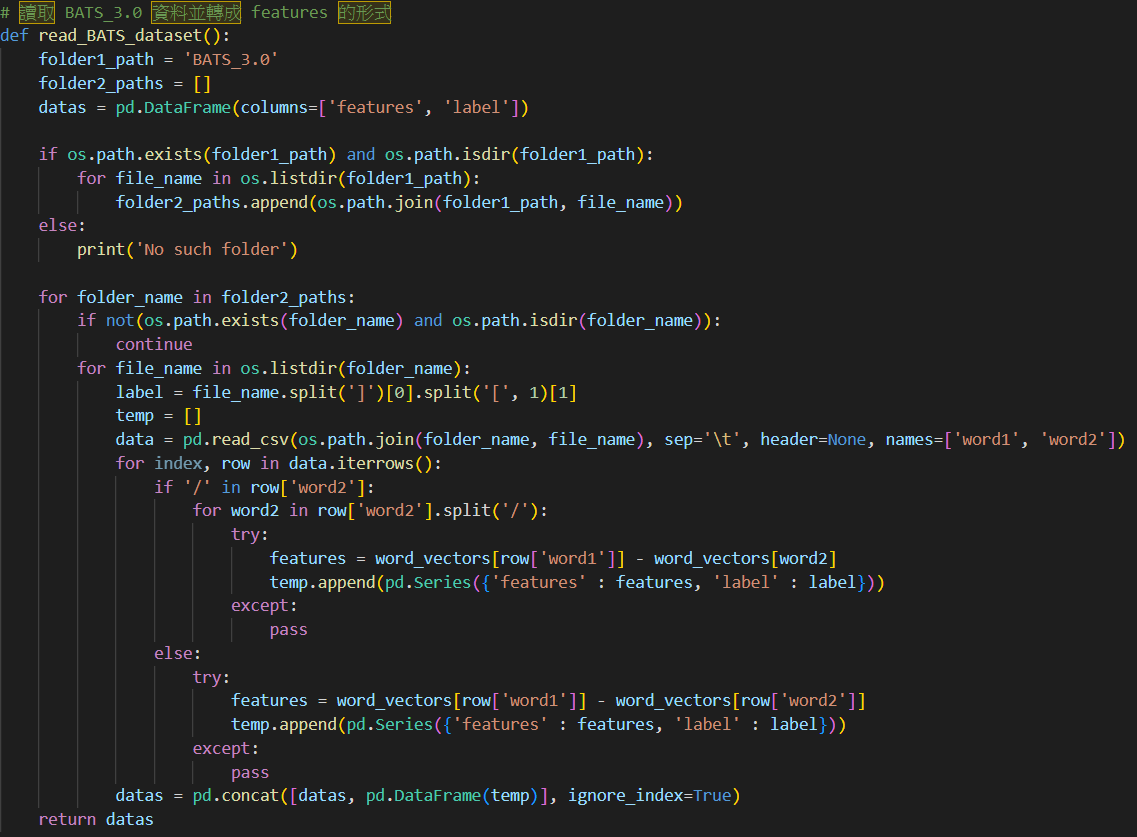


計算WordSim-353資料集各個單詞的相似度

輸出 correlation

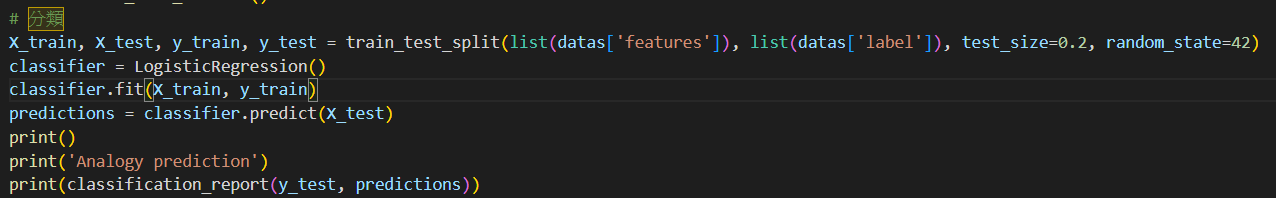


讀取 BATS\_3.0 資料並轉成 features 的形式



前段部分是在進入 BATS\_3.0 資料夾內層中的各個資料夾，之後將資料夾內的兩個或多個單字做向量的相減，並用此數值作為特徵值和資料夾名稱作為Label。

分類和輸出



用LogisticRegression進行訓練，前80%訓練，後20%測試。

執行結果：

