1. Data preparation: 將原本 json 檔的資料轉為 DataFrame 並儲存成 pickle 檔。
2. Data preprocessing:
   （1） 載入先前處理完的 DataFame
   （2） 刪除 duplicate 的 text
3. Model
   （1） 訓練 Word2Vec 模型
   （2） 使用 LightGBM 做分類任務
4. 預測 test_df 中的資料並產生 submission 的檔案
5. 結果與討論
   （1） 1. 一開始我使用 randomforest 配合 Tf-IDF 做為 baseline，Public 的準確度大概在 0.31 左右。
   （2） 2. 後來利用 Word2Vec 的值當做 feature，並用 LightGBM 做分類任務，因為數據集較大，用 LightGBM 比較快，Public 的準確度大概在 0.34 左右。
   （3） 3. 調整 Word2Vec 的參數，像是 vector_size, min_count, epochs 等等，雖然訓練時間拉長了一些，但是 Public 的準確度達到了 0.407。
   （4） 4. 後來嘗試使用 Bert，不過因為資料太大了，需要訓練的時間太長，所以沒能成功。