應用文字探勘分析新冠肺炎應對政策之網路聲量

第十八組程式系統執行說明

**Step1.程式環境安裝並匯入程式檔**

* 安裝Eclipse。

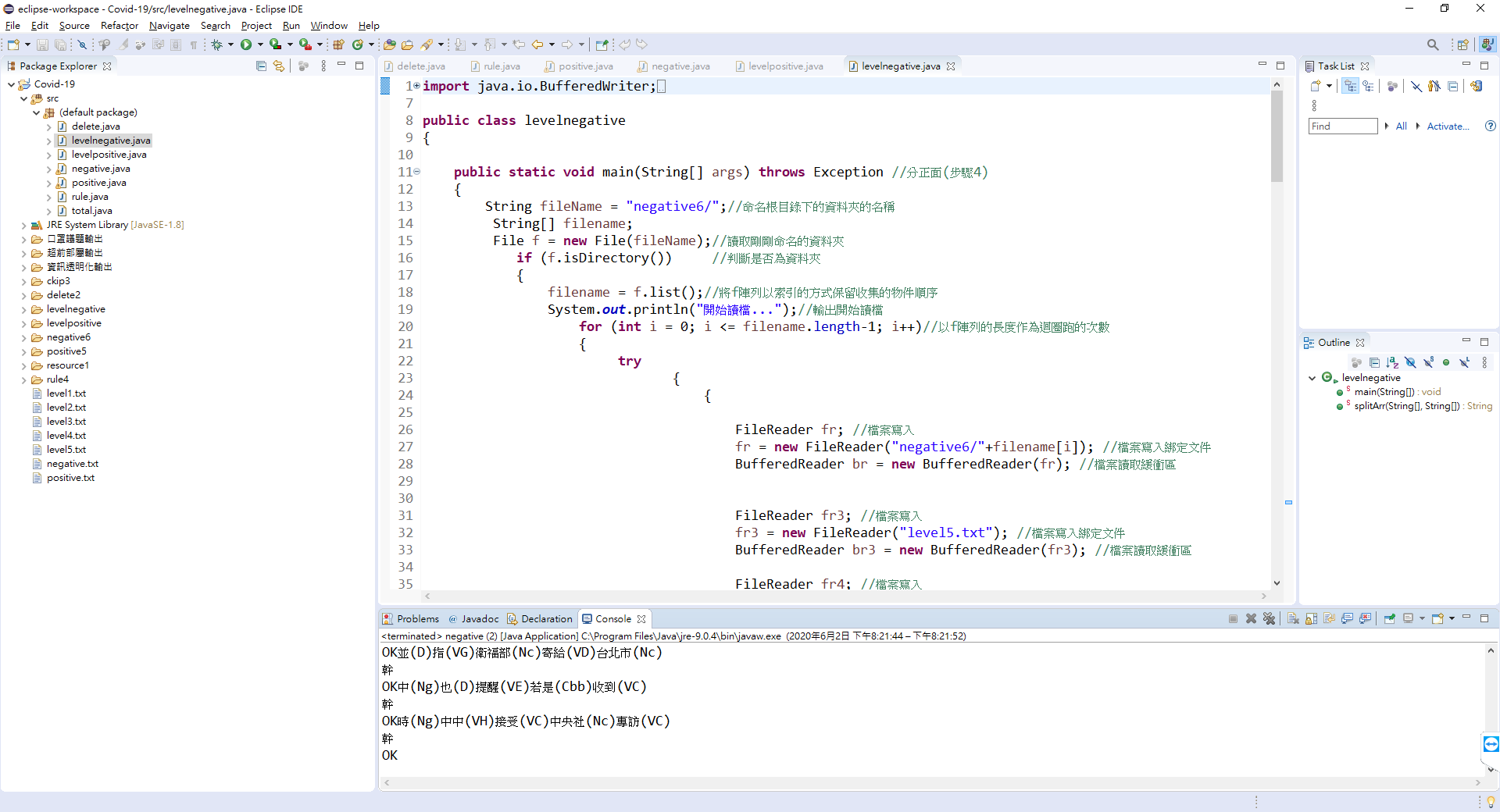
下載網址：https://eclipse.org/downloads/

* 安裝JRE與JDK，並設定JAVA開發環境。

下載 Java Development Kit (JDK)網址：http://www.oracle.com/index.html

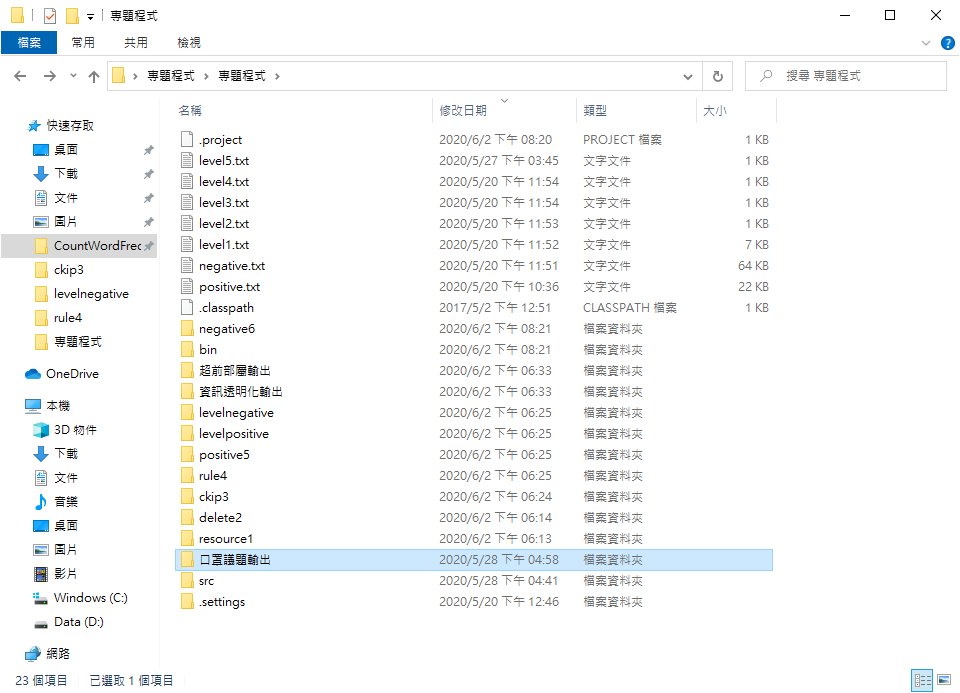
* 匯入程式檔，如下圖。
* 下載CKIP，需自行申請帳號才可獲取解壓縮密碼。

網址：http://ckipsvr.iis.sinica.edu.tw/ckipws/reg.php



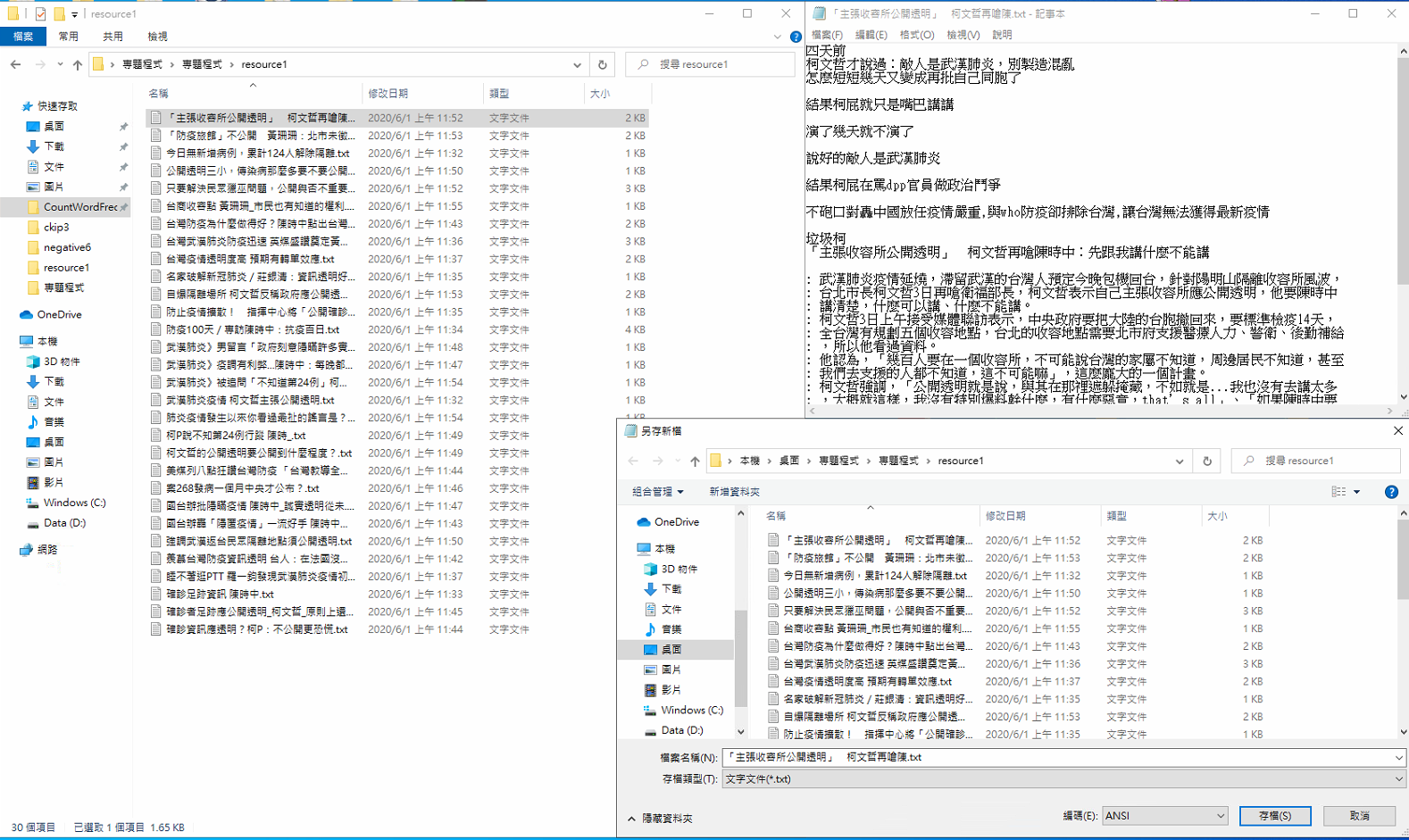
**Step2.先建立以下幾個資料夾：**

* resource1、delete2、ckip3、rule4、positive5、negative6、levelpositive、levelnegative



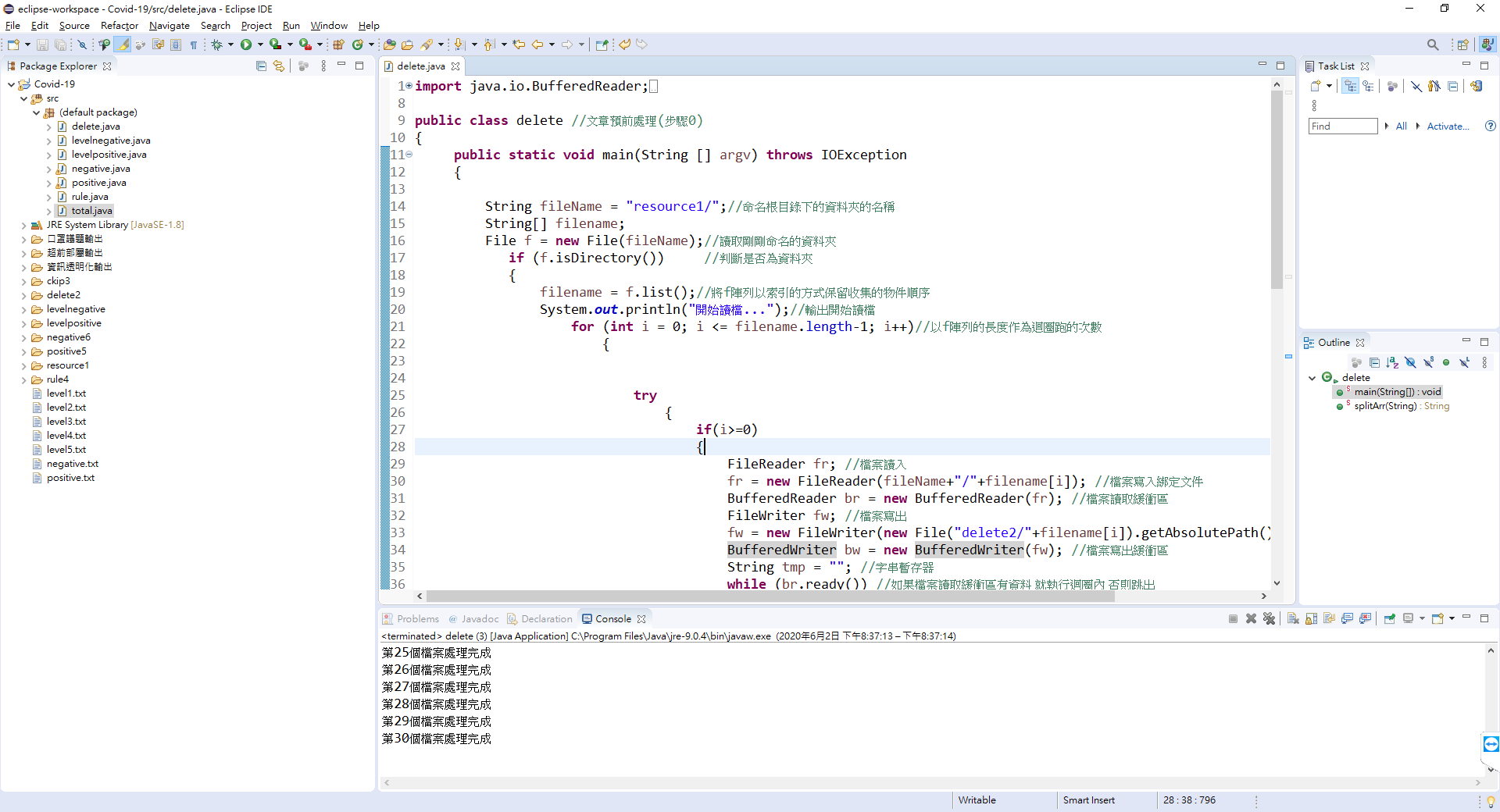
**Step3.資料彙整**

* 將收集來的資料整理成TXT檔並存入resource1，記得將記事本另存新檔為ANSI編碼

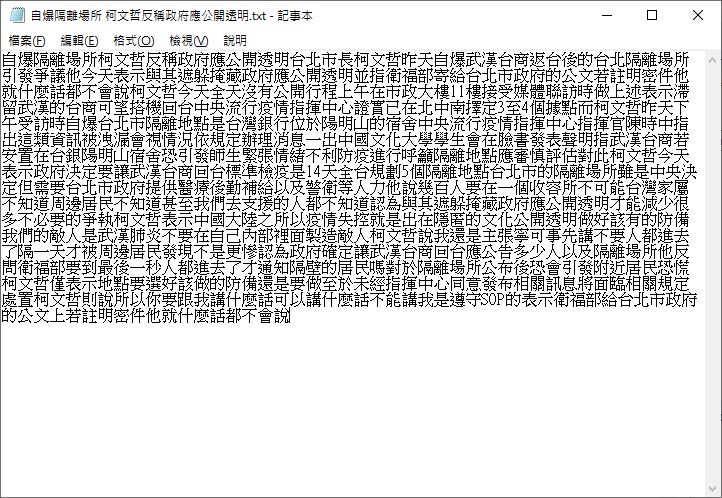


**Step4.執行程式delete.java**

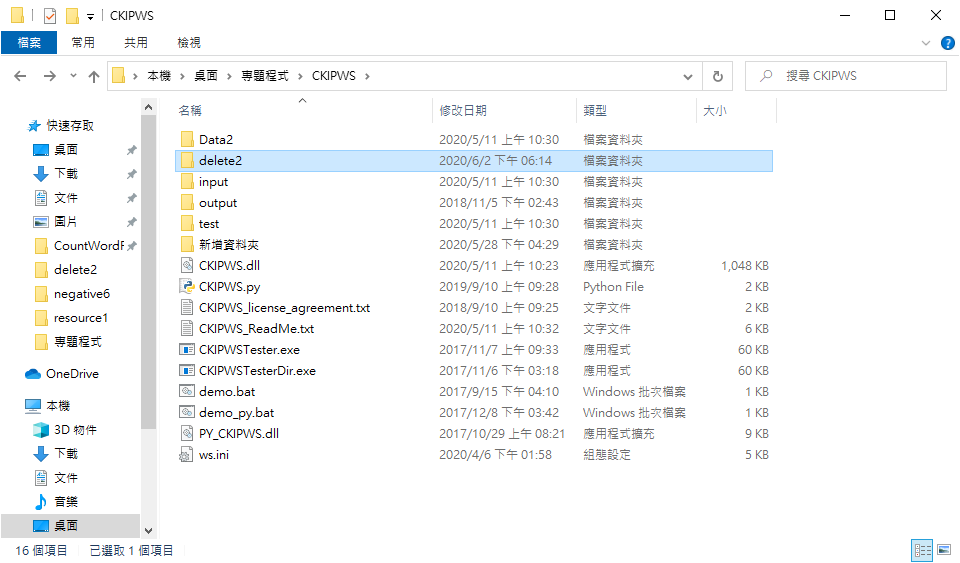
* 執行程式，會將resource1內的文字檔標點符號全部刪除。



* 結果輸出如下

**Step5.執行CKIP**

* 將delete2資料夾複製到CKIPWS資料夾下

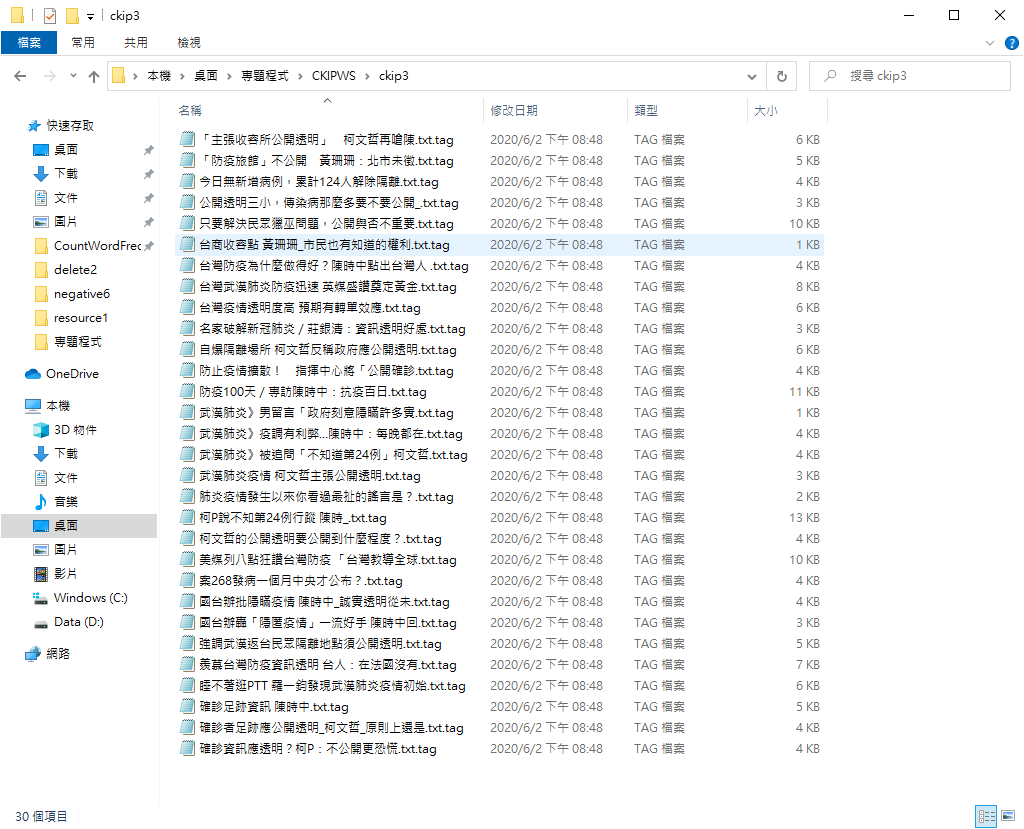


* 並在開始->執行->輸入CMD並且CD進入CKIPWS資料夾後輸入指令，詳細指令請看CKIPWS\_ReadMe.txt可能隨版本不同指令會有所不同。

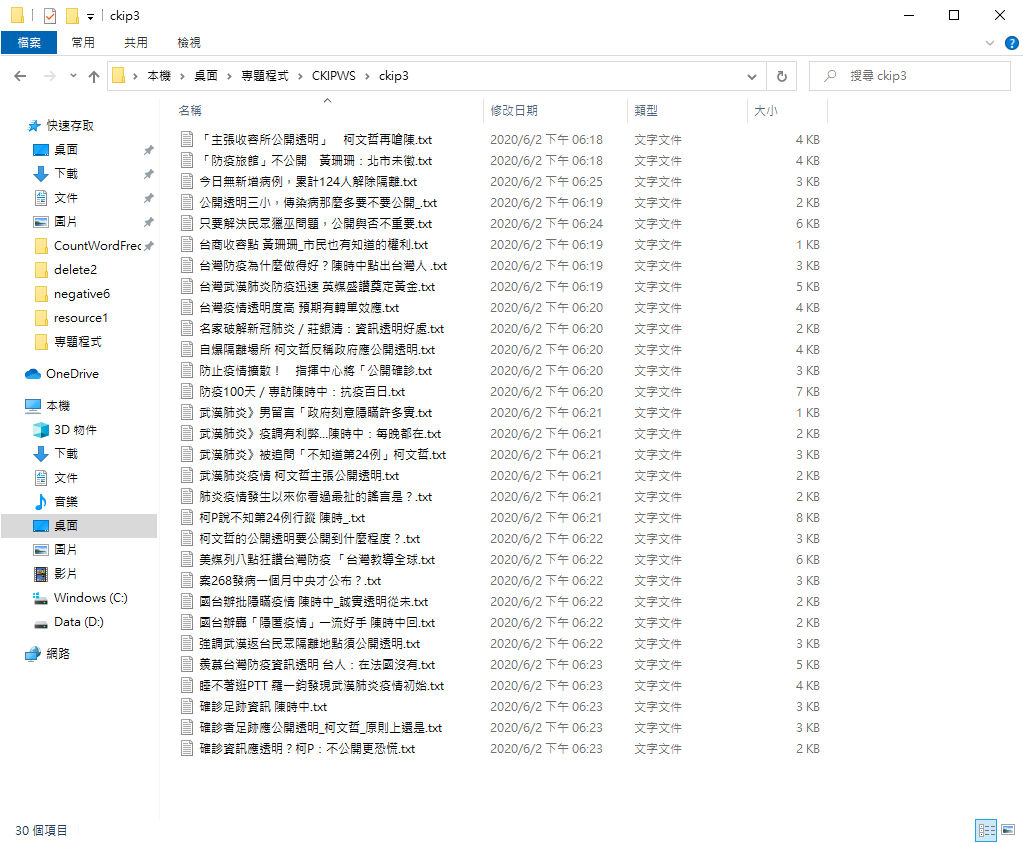
註：CKIPWS時常會推出新版本，只要新版本一出現舊版本馬上就不能用，所以請保持CKIPWS版本是最新的。



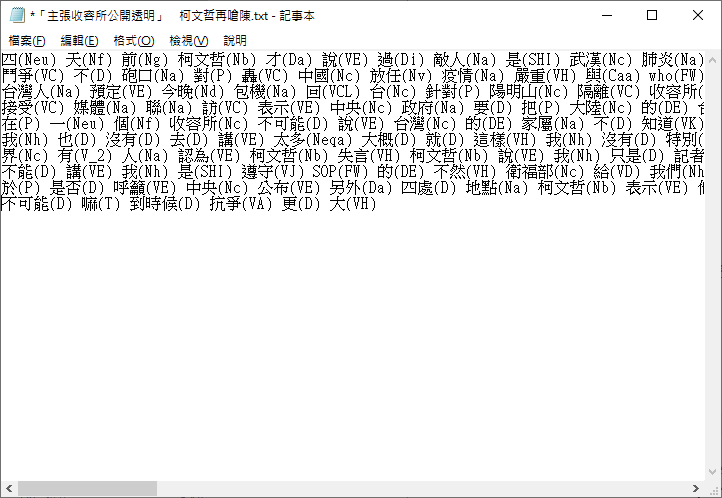
* CKIP執行後會產生出處理後的資料



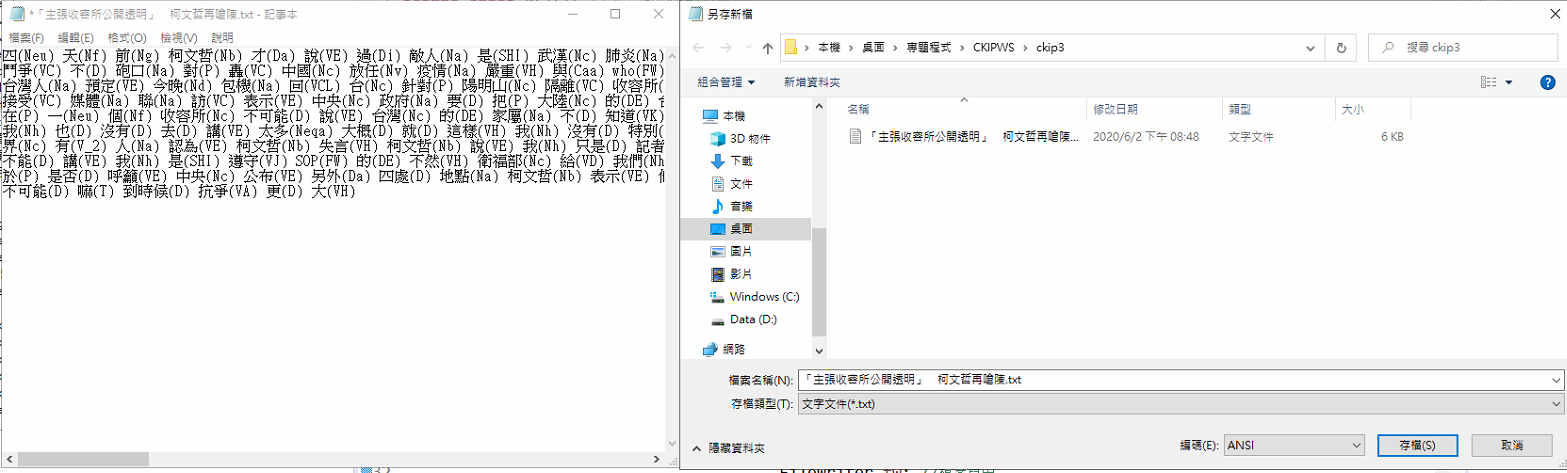
* 這裡必須將後面的.tag刪除，處理後如下。



* 並將裡面所有全型空白符號全部取代為半型空白符號，並且將每一列的第一個空白符號刪除，處理完結果如下。



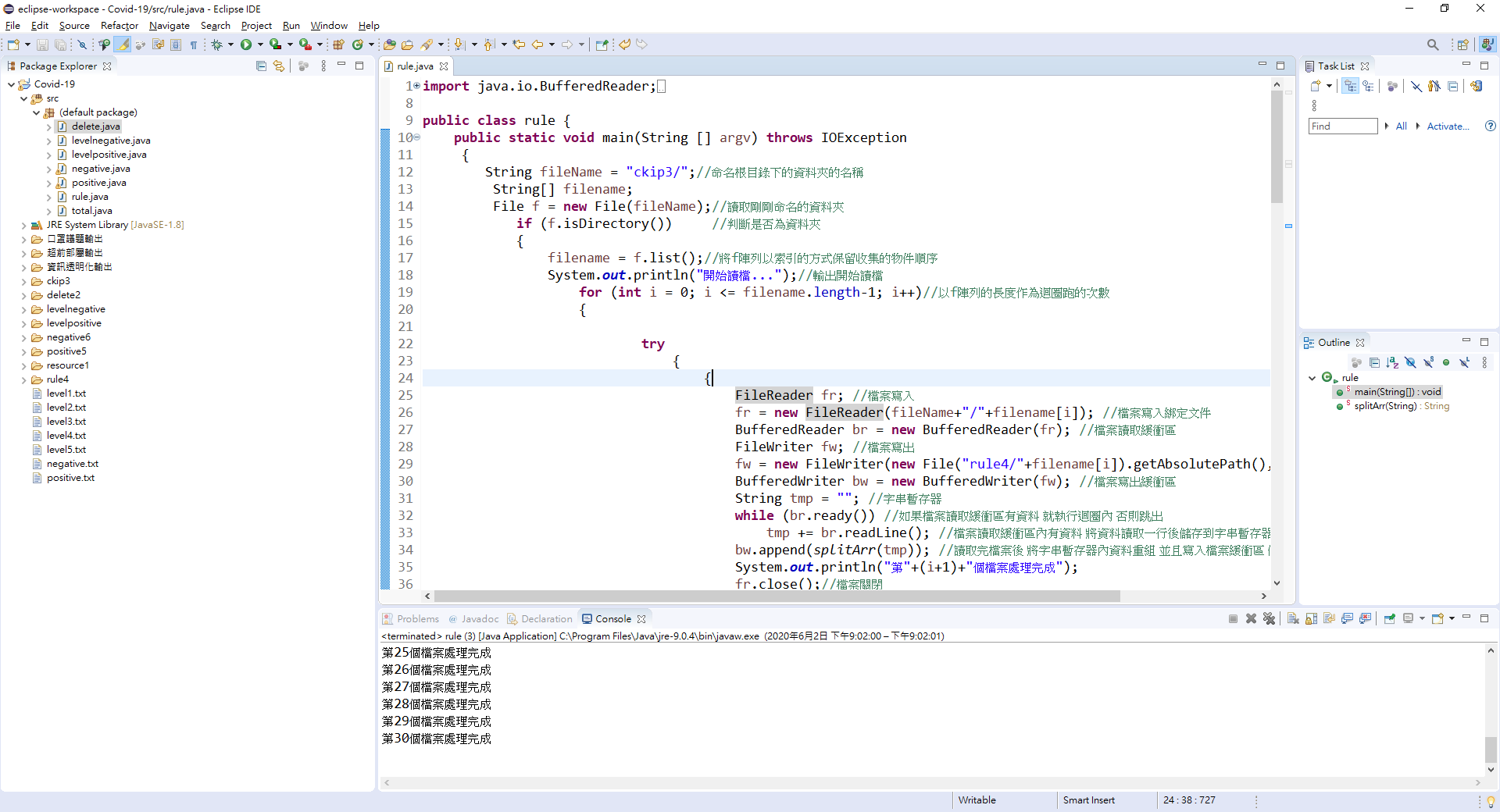
* 因執行CKIP輸出結果會是UTF-16LE，所以必須另存新檔確保編碼為ANSI。



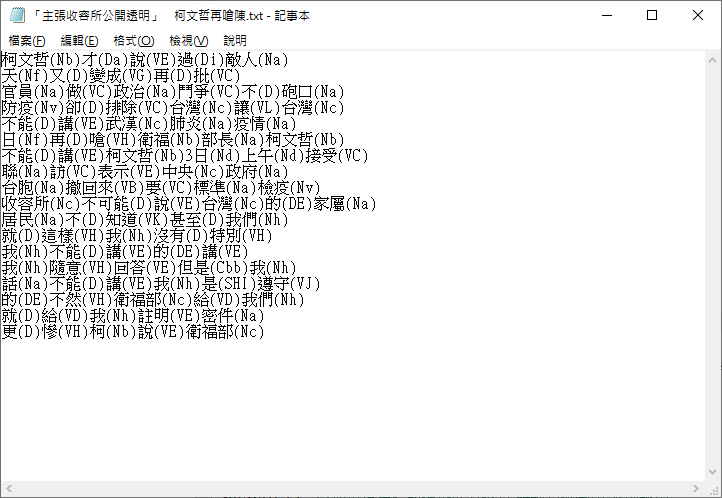
* 處理完後將ckip3資料夾移置原程式執行的資料夾。

**Step6.執行程式rule.java**

* 執行程式，會輸出詞組規則。

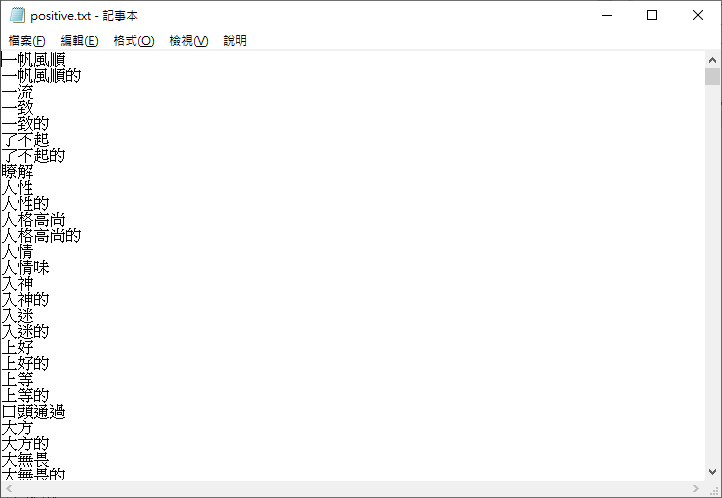


* 結果輸出如下



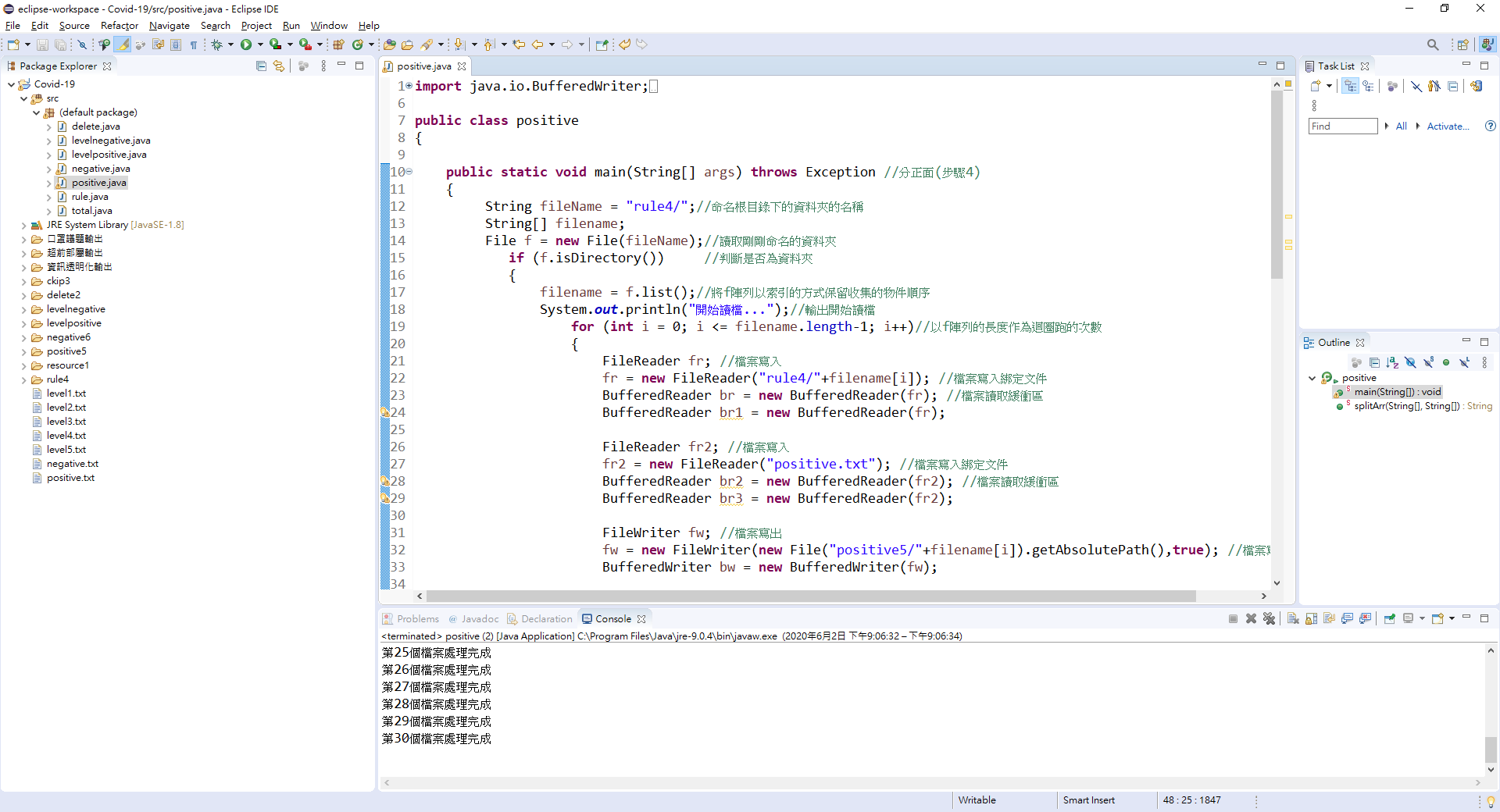
**Step7.蒐集正面詞、負面詞、程度詞詞庫**

* 先到網路上尋找相關的詞庫，並且建立以下幾個TXT檔，分別為positive.txt、negative.txt、level1、level2、level3、level4、level5並將每個詞以換行符號做區隔，如下圖。

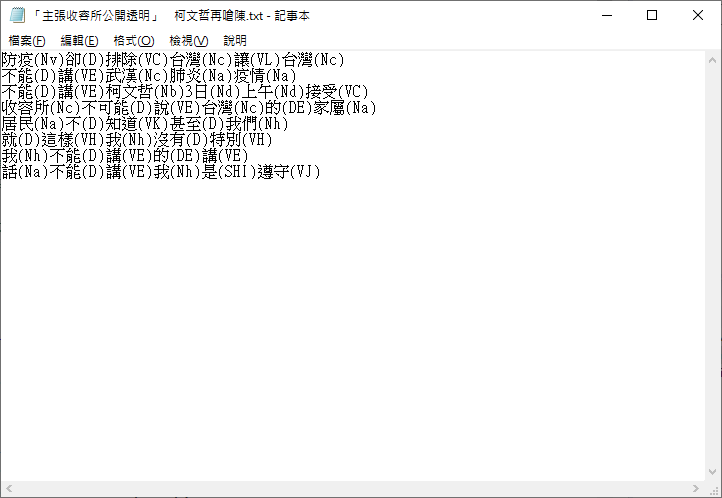


**Step8.執行程式positive.java**

* 執行程式，會根據詞組規則輸出相關的正面詞語。

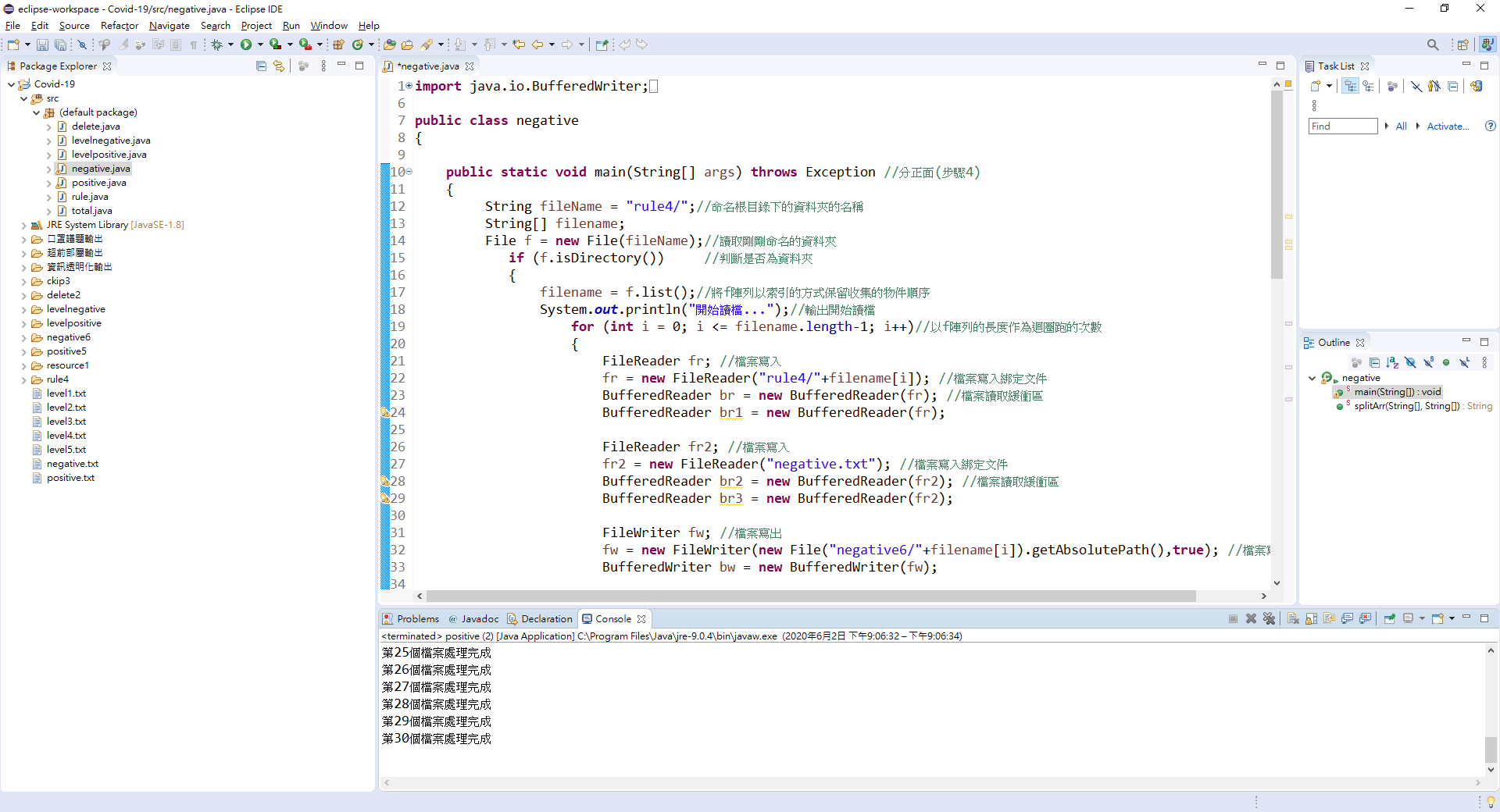


* 結果輸出如下

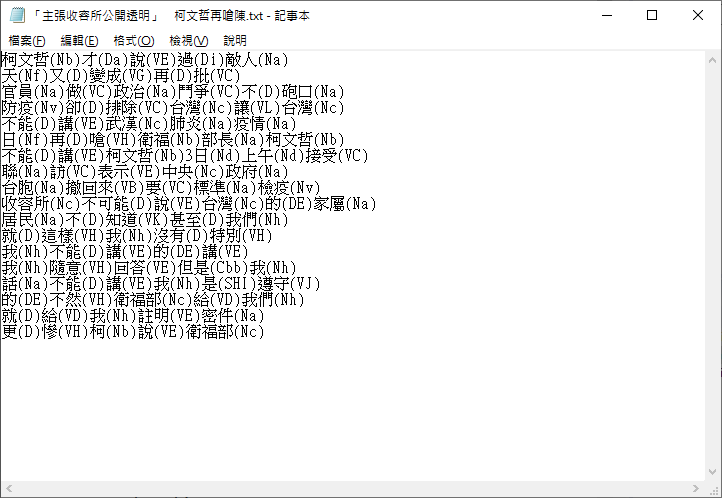


**Step9.執行程式negative.java**

* 執行程式，會根據詞組規則輸出相關的負面詞語。

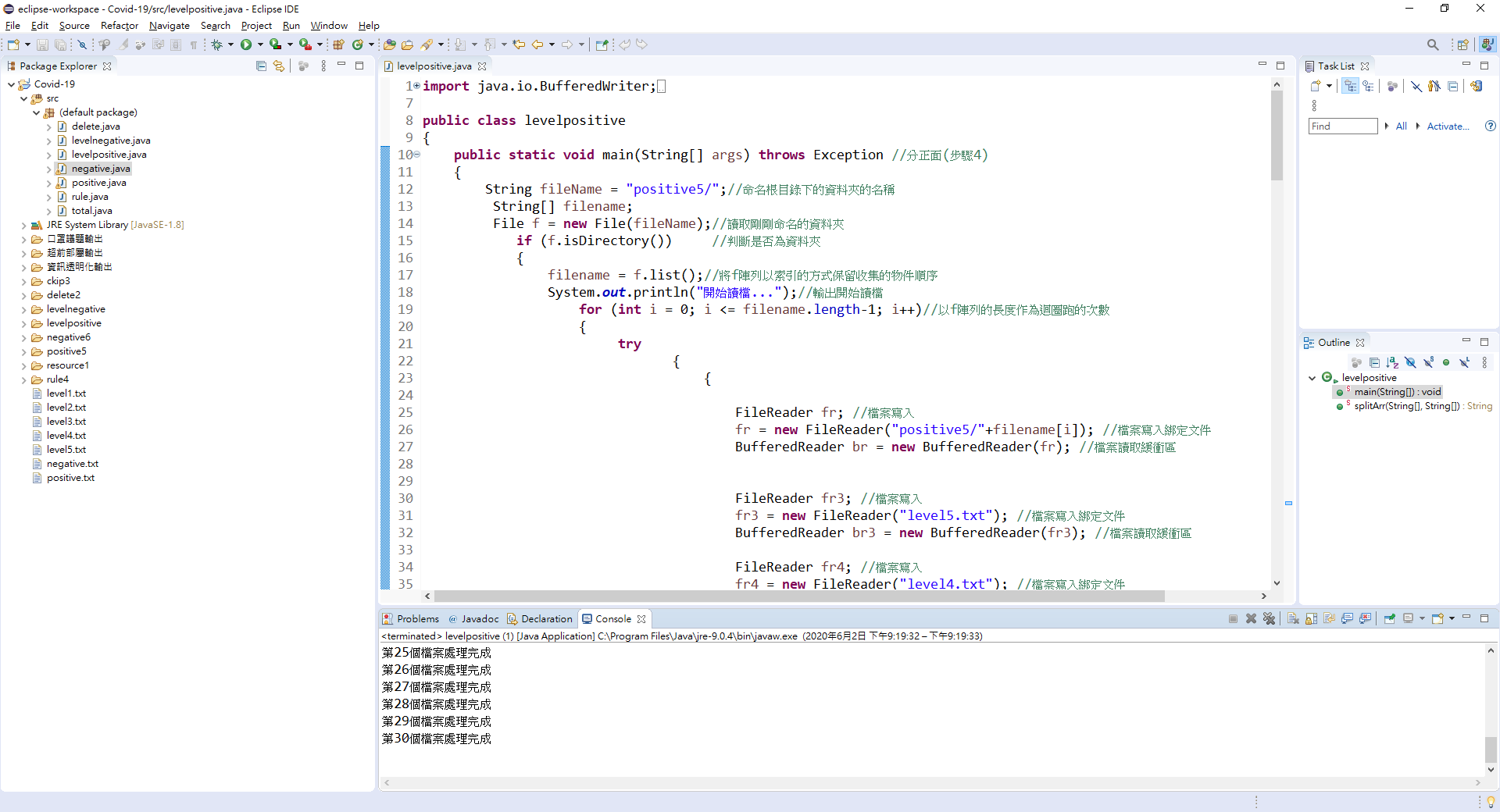


* 結果輸出如下

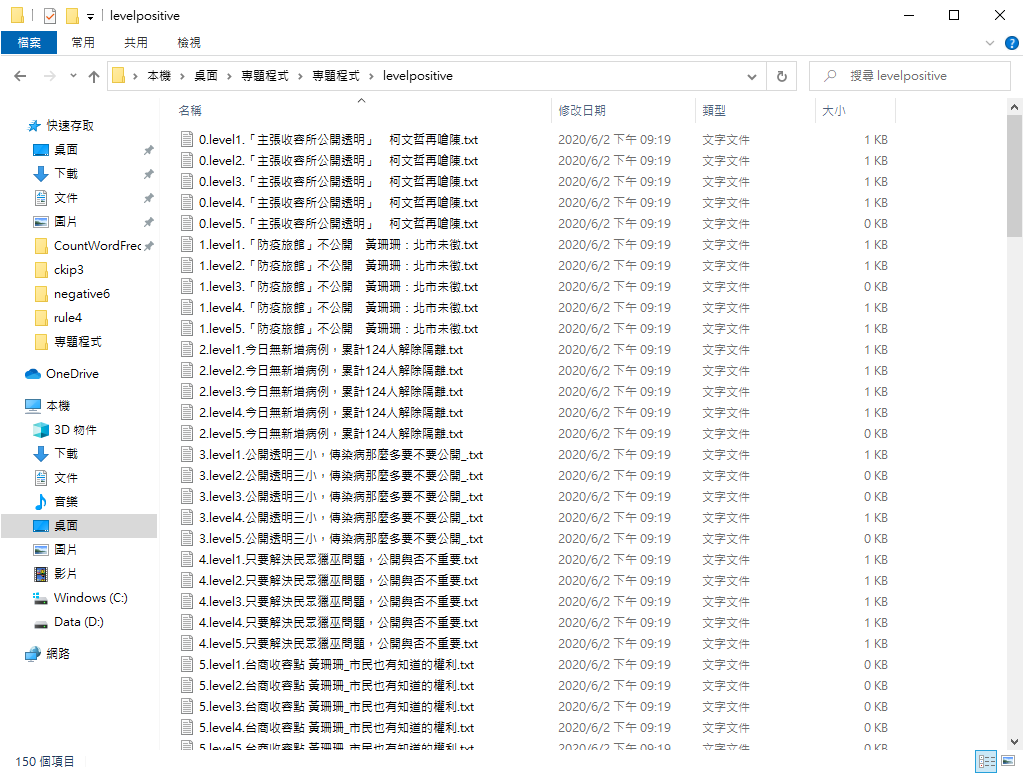


**step10.執行程式levelpositive.java**

* 執行程式，會根據步驟8所產生出的正面詞與輸出五個階級的正面程度詞。

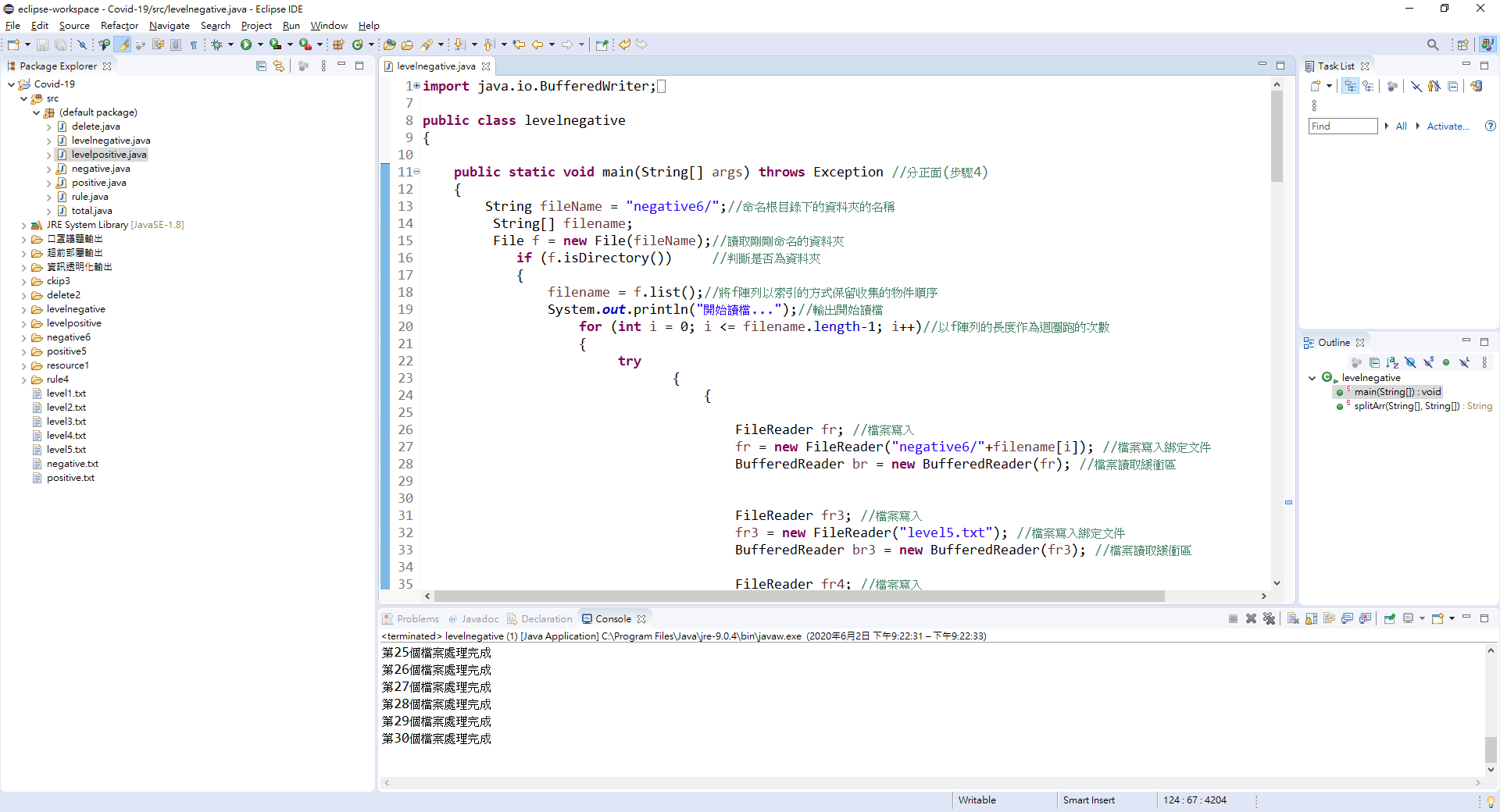


* 結果輸出如下

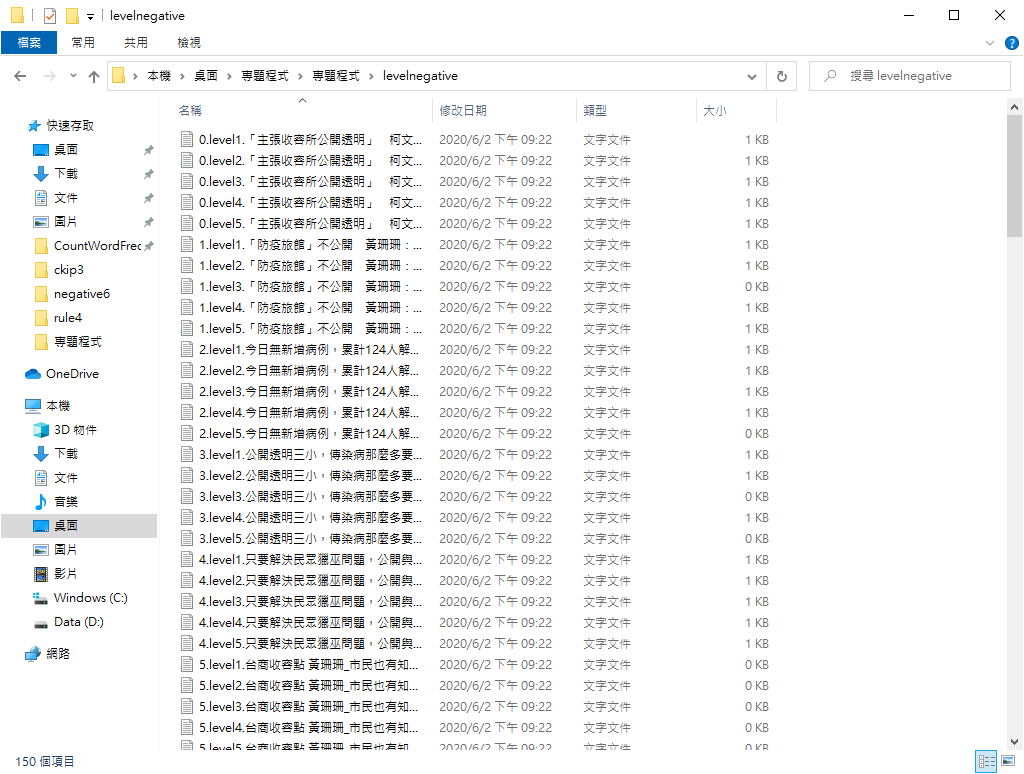


**step11. 執行程式levelnegative.java**

* 執行程式，會根據步驟9所產生出的負面詞與輸出相五個階級的負面程度詞。



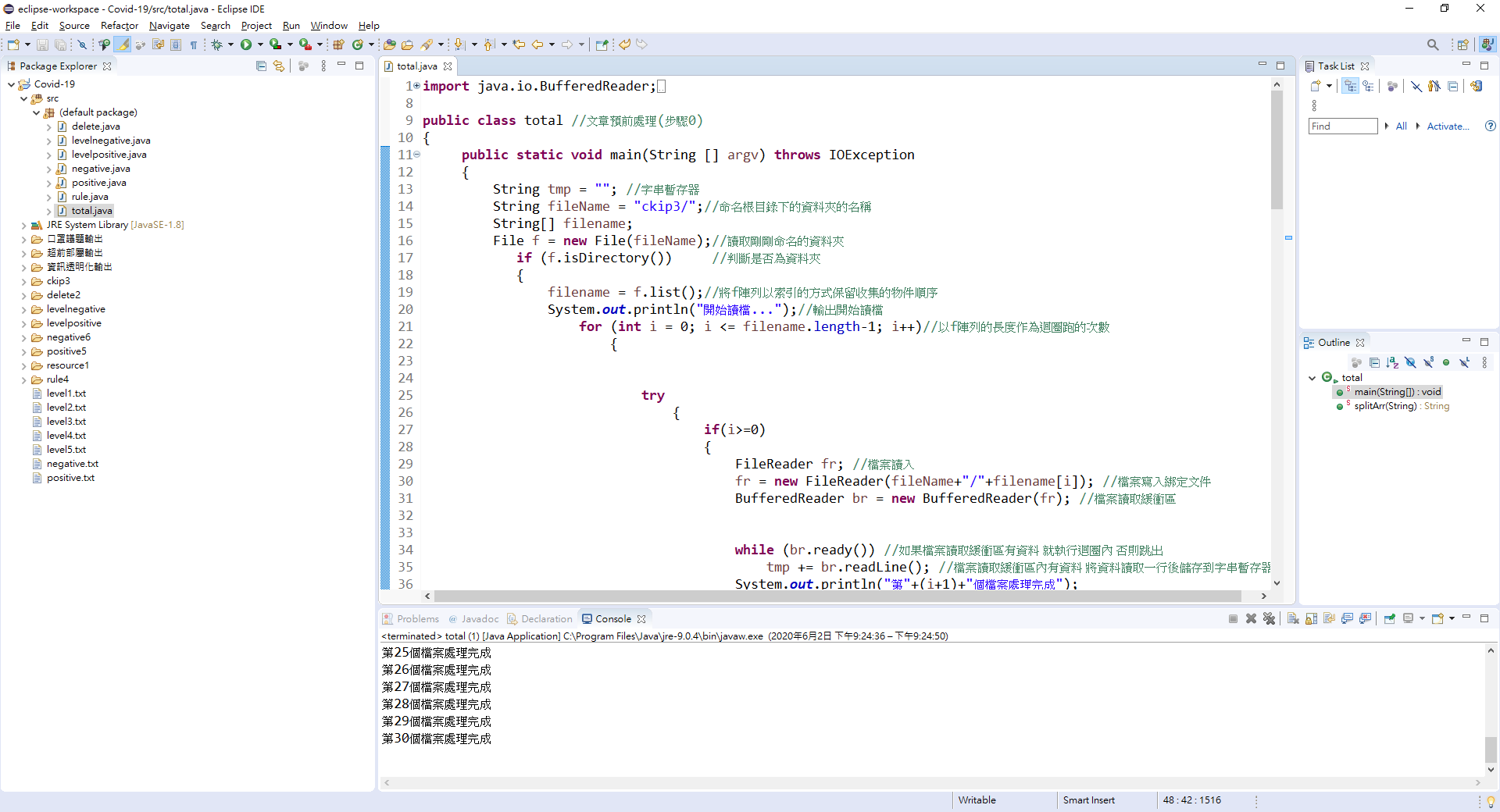
* 結果輸出如下



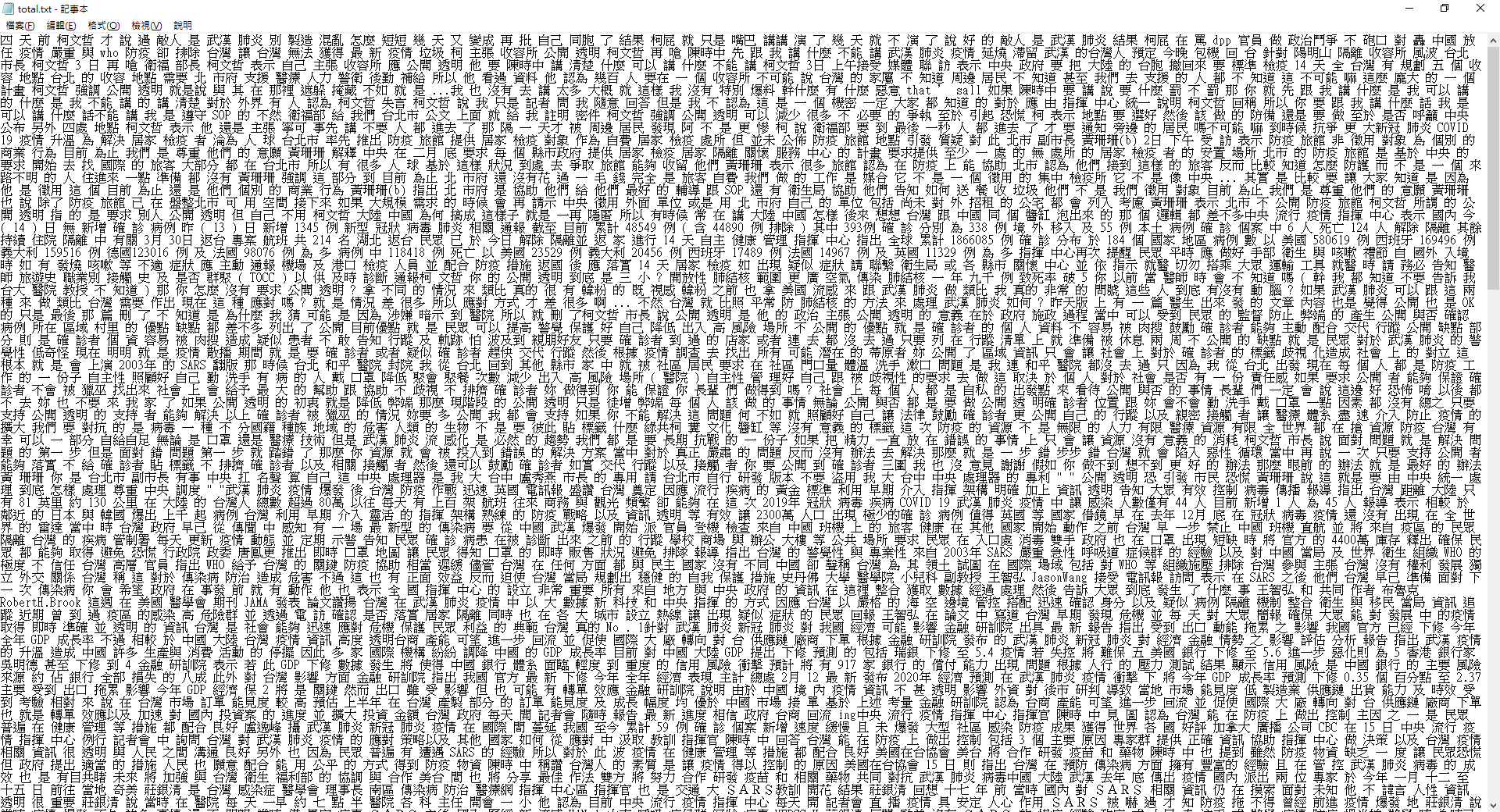
**Step12.執行程式total.java**

* 執行程式可將所有文章匯集在同一個記事本，方便後面計算詞頻。

註：因作業量龐大，因此執行可能會需要一些時間



* 執行完後會產生total.txt

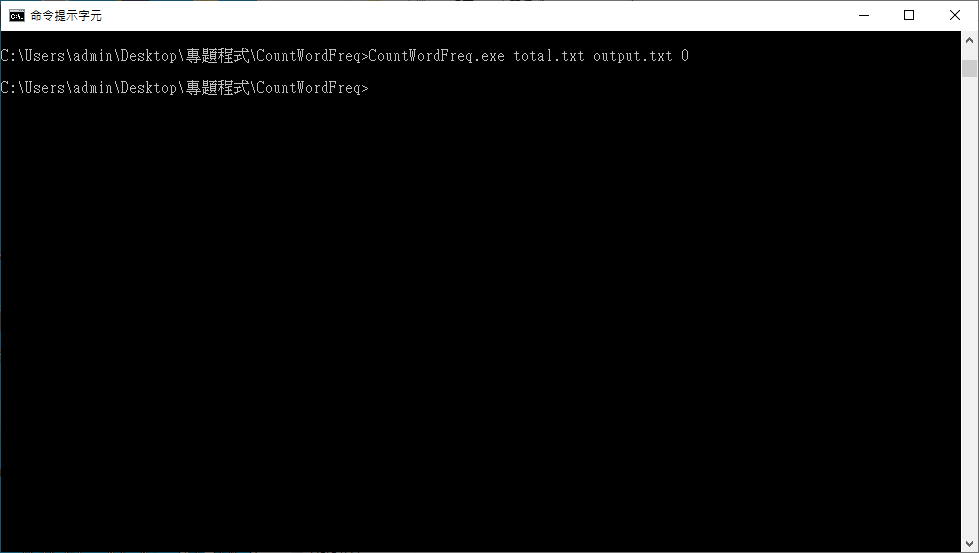


**Step12.計算詞頻**

* 請先至http://ckipsvr.iis.sinica.edu.tw/resource.htm下載CountWordFreq詞頻計算系統，並且將上述的total.txt內的半型空白符號全部取代為TAB符號並存檔於CountWordFreq資料夾下，取代完結果如下。



* 在開始->執行->輸入CMD並且CD進入CountWordFreq資料夾後輸入指令，詳細指令請參考上述網站。



* 結果如下



* 將上面執行後所產生的output.txt於Excel 資料->從文字檔->選取output.txt->下一步->下一步->完成->確定。結果如下

