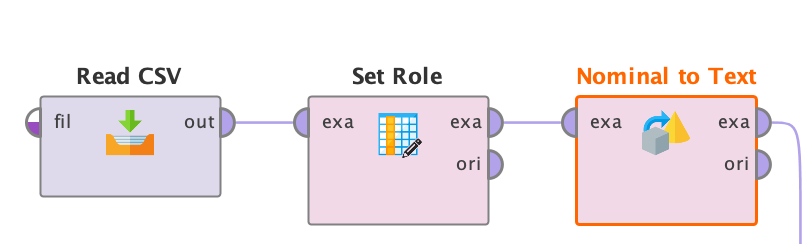
作業4: 眾裡尋他千百度,那模型就在燈火欄珊處

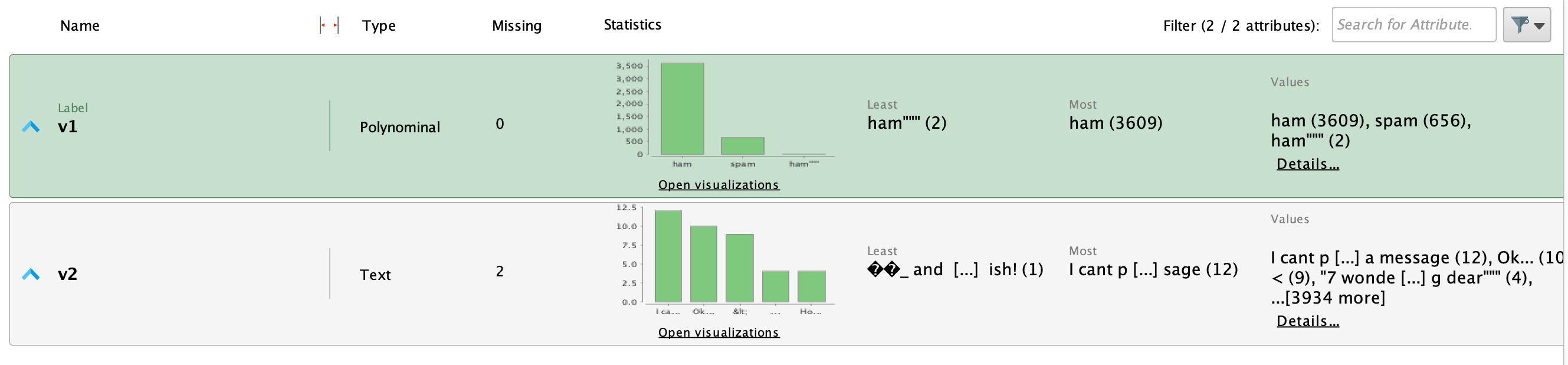
現在我們就要開始來使用江湖中傳說的文字探勘絕學啦!

為了確保大家的學習品質,因此課程中一半的內容將做為作業的一部份;另一部份才是讓大家帶回去練習的作業唷!那我們就開始吧!!

1. 請問啟spam.xls檔案! 並開始進行分析,開始前請針對需要的欄位進行必要之前置處理. (10%)

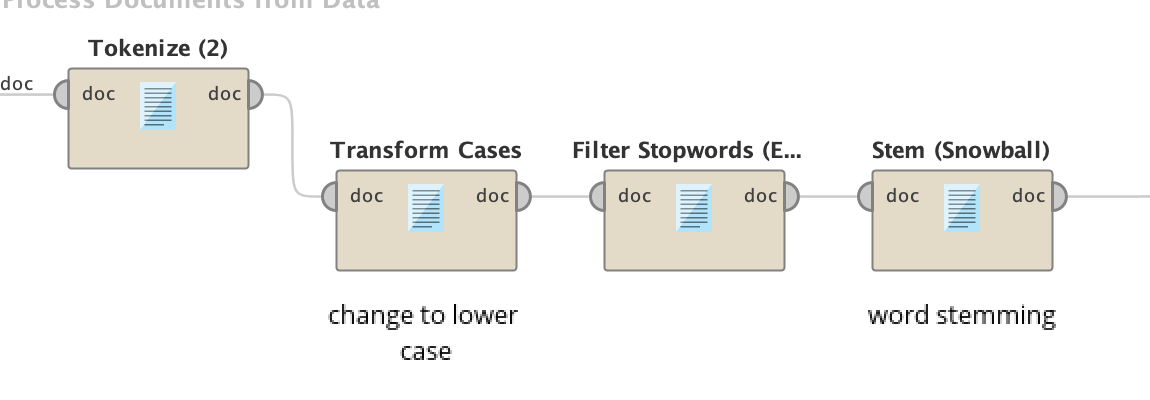
檔案讀取後，先進行label的設定，把v1欄位當作label，接著為了處理文字，使用類別轉文字，並設定要處理的欄位是v2。



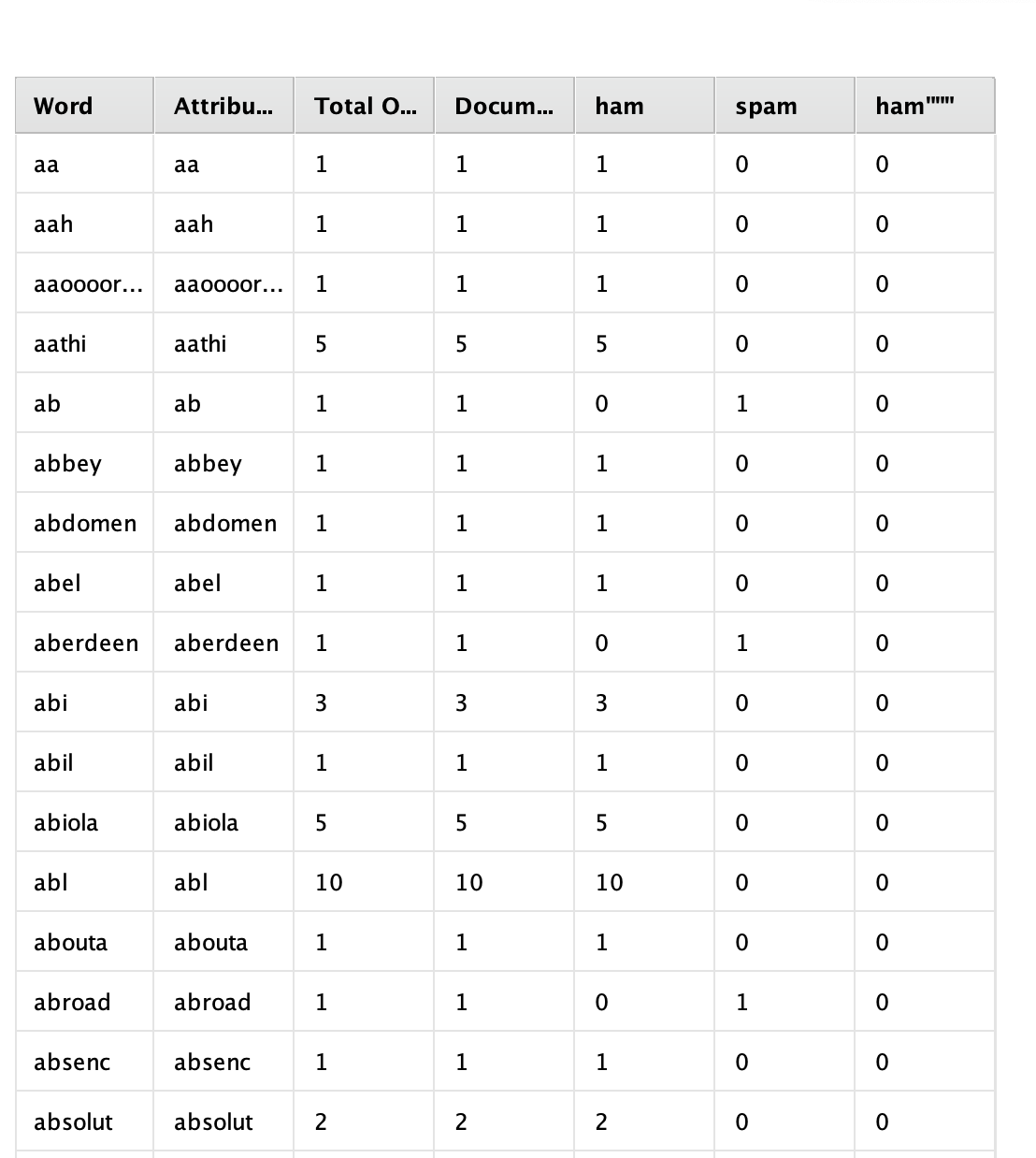


2. 接下來,我們要針對文字進行分析,因此有一些必要的處理程序,請說明一下您的步驟(10%),並請說明下列分析結果的意涵: TF-IDF (5%)

開始使用text processing，先斷詞、將全部的字轉成小寫、過濾停用字，最後取詞幹，將字型還原。

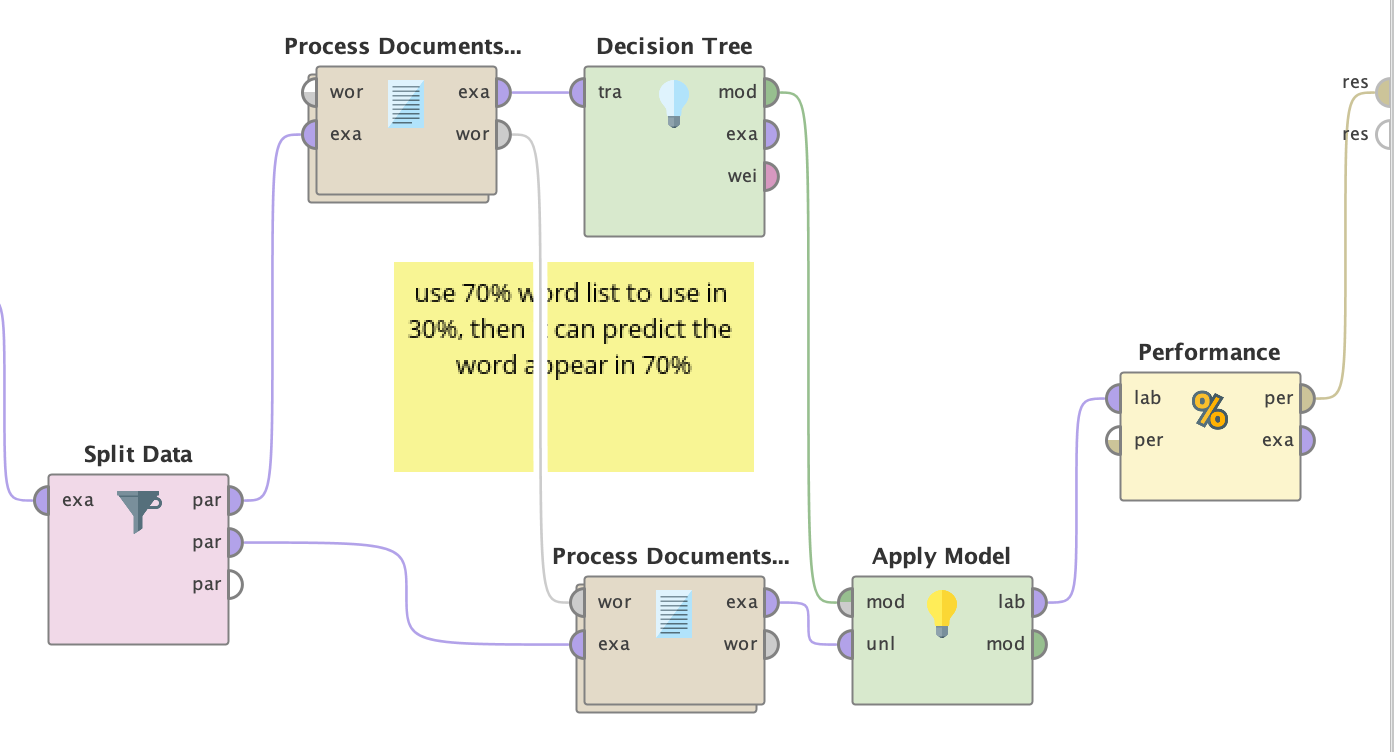


TFIDF是透過字詞出現的頻率來計算，以下圖片顯示每個字在檔案內出現的次數。

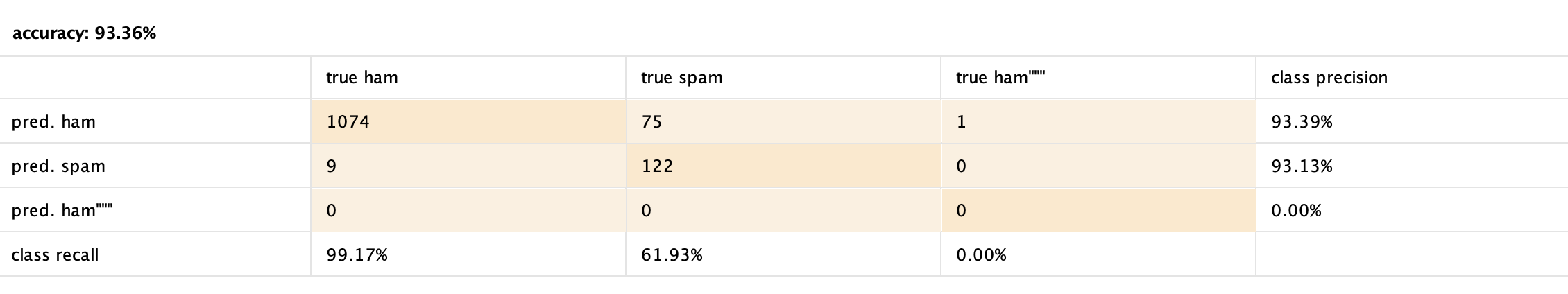


3. 接下來,我們要開發製做一個垃圾郵件過濾器,請找出您能找出正確率最高的分類機制,並略加說明(15%)

\*註:因為有正確率評估的關係,請記得要使用訓練/測試集,又或是Cross-Validation



為了進行訓練及測試，先將數據切分成70%訓練30%測試，然後訓練集文字預處理後，把word list丟入測試集的文字預處理，即可開始模型預測，本次模型選用決策樹。



模型預測結果準確度為93.36%，其中，1074個是一般信，預測結果也為一般信，9個是一般信但卻被預測為垃圾信，75個垃圾信被預測為一般信，122個垃圾信且被預測為垃圾信。

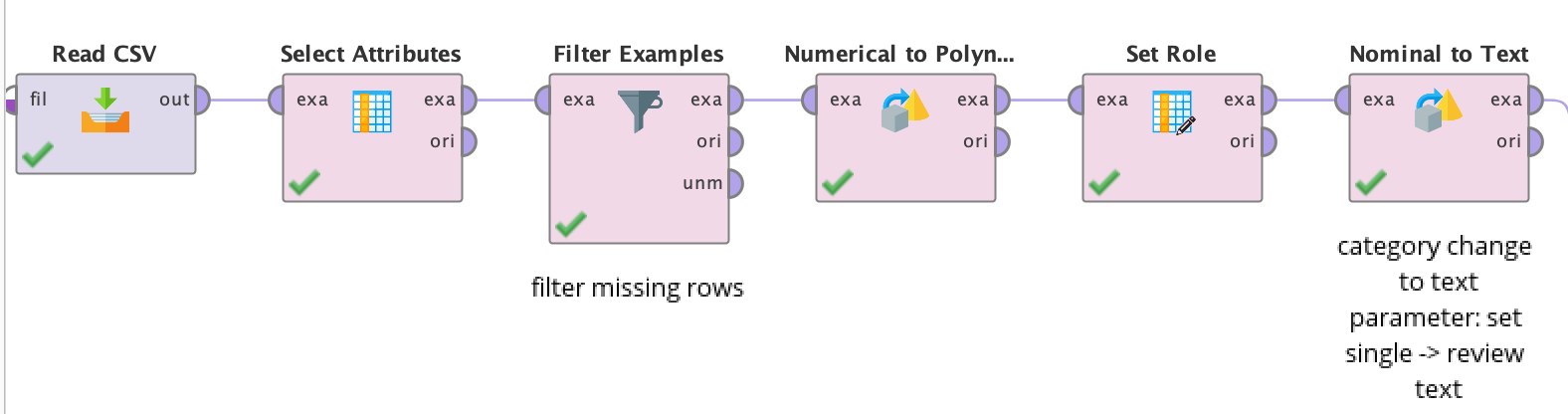
4. 接下來,承上述步驟,請使用Womens Clothing E-Commerce Reviews.csv檔案進行分析,在這邊我們希望您可以借由分析評論來預測評分或是您認為值得分析的議題. 另外也請您自行思考一下還有什麼分析可以進行並說明結果唷 (60%)。

其他可分析議題：

可以使用評論來分析顧客的回饋，並探討哪個服飾類別的整體回饋表現出較多負面的評價，讓電商可以針對該類別進行改進的動作。

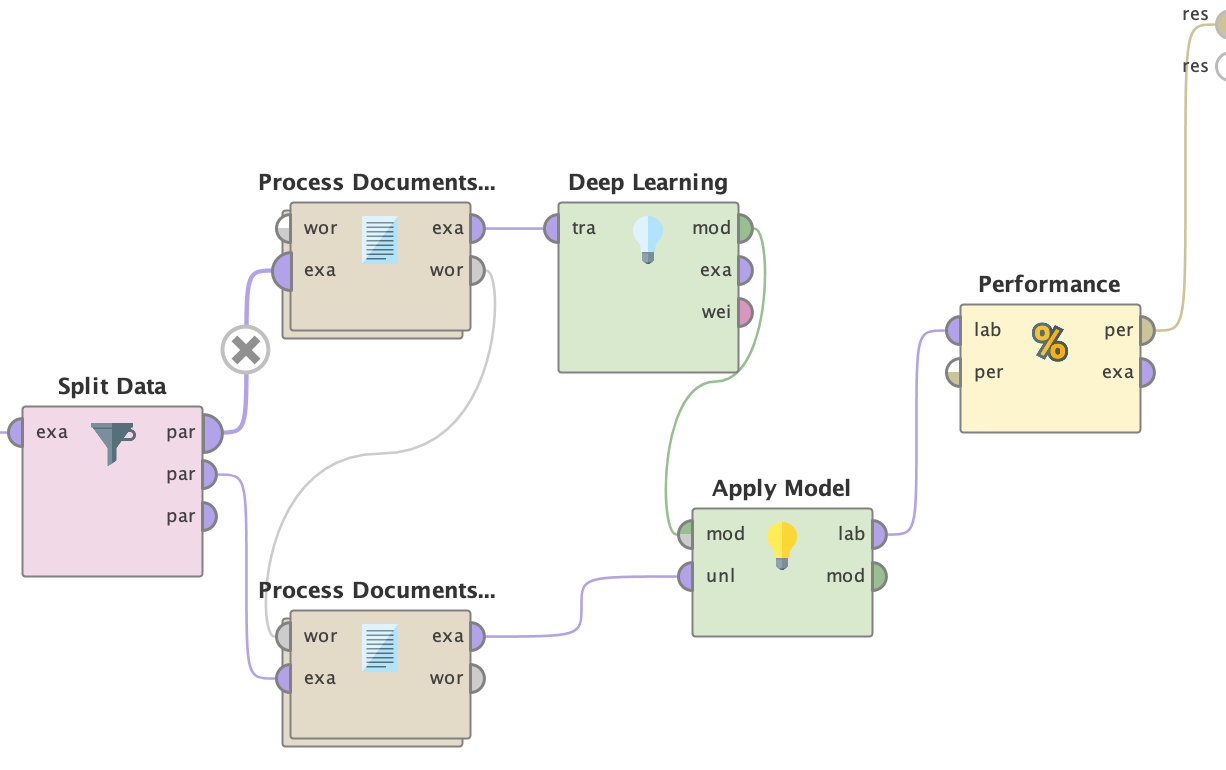
以下使用僅使用評論來預測評分。

**文字處理前的前處理：**

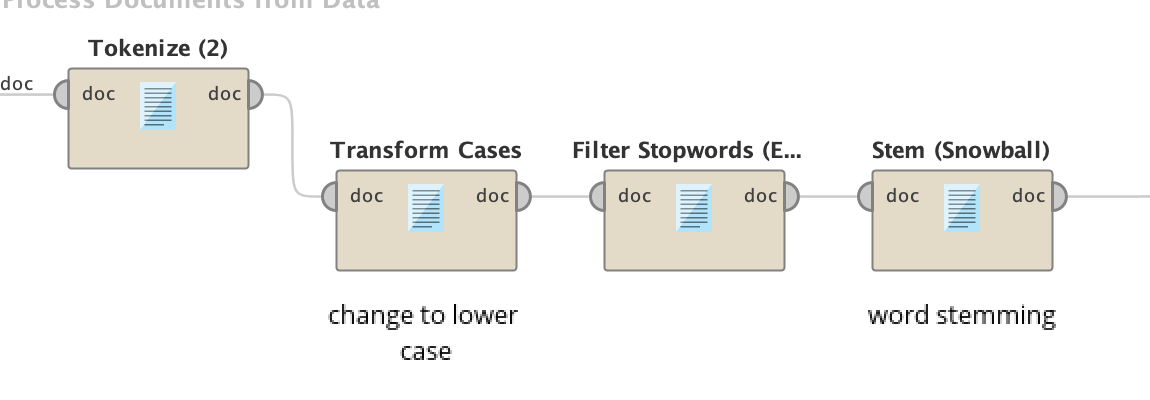
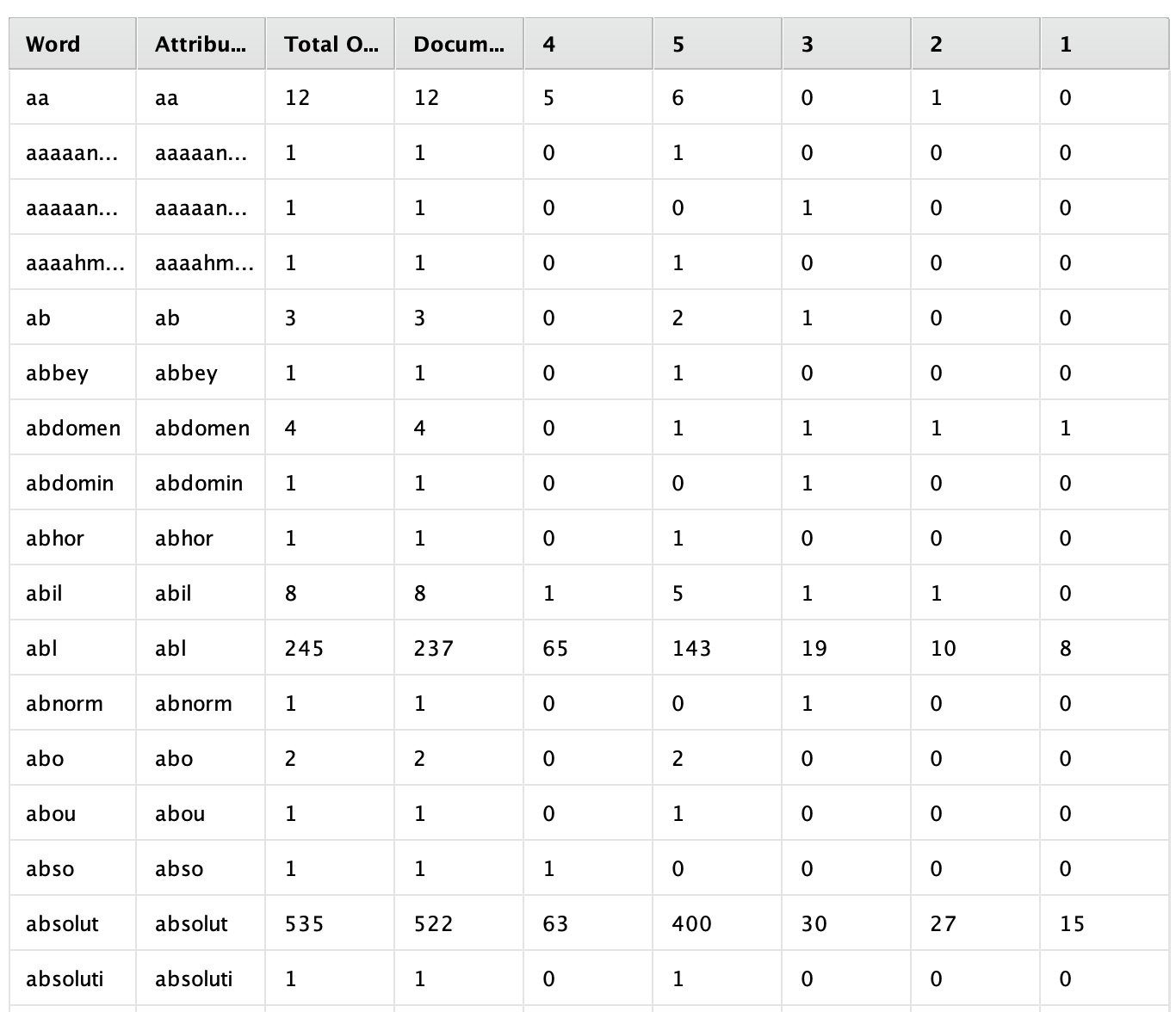


讀取檔案後，先取得要使用的欄位，本次選取review text、rating兩個欄位。接著針對？值進行資料過濾。為了讓模型能預測rating欄位，將rating 欄位的型態轉為文字類型，並將它設為label。然後類別轉文字，並設定要處理的欄位是review text。

**訓練及測試、模型預測：**



**文字預處理：**

將訓練集與測試集切分成70%與30%，然後分別進行文字預處理(包含斷詞、全部的字轉成小寫、過濾停用字，取詞幹)，此時，訓練集文字預處理後需把word list丟入測試集內進行文字預處理，最後才開始模型預測。

**模型結果：**

模型這邊選用深度學習，結果可如下圖看到準確度為57.16%，可以發現評分5分且預測結果5分的數量為2884。

