作業6: 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%)

〔前處理〕

一張含有 文字, 螢幕擷取畫面, 圖表 的圖片

自動產生的描述

* 1. Retrieve spam\_clean：讀取spam\_clean.csv資料
  2. Set Role：設定v1欄位為label  
     一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

     自動產生的描述
  3. Nominal to Text：資料轉為text型式  
     一張含有 文字, 螢幕擷取畫面, 字型, 陳列 的圖片

     自動產生的描述  
     ※ 由於讀入的資料其v2欄位屬性為object，因此要轉text  
     ※ 使用Nominal to Text 將v2轉換為text
  4. Split Data：切分資料，訓練:驗證=70%:30%  
     一張含有 文字, 螢幕擷取畫面, 軟體, 陳列 的圖片

     自動產生的描述
  5. Replace：其中有兩筆資料讀成ham"""，改為ham。

一張含有 文字, 螢幕擷取畫面, 陳列, 數字 的圖片

自動產生的描述

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%)

1. Step 1.

*Import text data with some assessment of the sentiment related to it. It is processed to extract the words and deliver a word-vector (a numerical representation of the text).*

* 資料前處理（見第一題描述）
* 導入文本數據，將文本拆解為Token。

一張含有 文字, 螢幕擷取畫面, 圖表 的圖片

自動產生的描述  
※ 使用Process Documents from Data 來進行文字的處理

一張含有 文字, 螢幕擷取畫面, 字型, 圖表 的圖片

自動產生的描述

* 使用Tokenize進行斷詞斷句
* 使用Transform Case進行大小寫轉換，這裡將大寫都轉換為小寫。
* 使用Filter Stopwords(English)將停用詞刪除。

1. Step 2.

*Train the model and validate it to collect the performance data.*

訓練模型

一張含有 文字, 螢幕擷取畫面, 圖表, 行 的圖片

自動產生的描述

一張含有 圖表, 文字, 行, 螢幕擷取畫面 的圖片

自動產生的描述

↑上圖，這裡以Decision Tree為例。

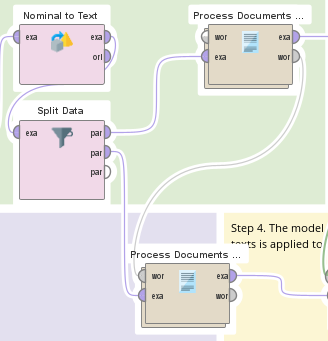
1. Step 3.

The model trained with the old texts is applied to the new document.

用訓練模型應用於Testing Data。

一張含有 文字, 螢幕擷取畫面, 圖表, 字型 的圖片

自動產生的描述

* Testing Data也要做斷詞斷句處理（Process Documents to Data），並且要給word，讓它知道如何斷句。
* 使用Apply Model將訓練完的模型拿去與Testing Data進行測試
* 接上Performance來看模型的結果。  
    
  ↑上圖，由於Training Data和Testing Data文字處理出來的字會不一樣，這將會造成後面模型建置產生問題，所以要透過將訓練集的word接出，連接至Testing Data的word。
* 整體Process圖

一張含有 文字, 圖表, 螢幕擷取畫面, 地圖 的圖片

自動產生的描述

* TF-IDF 一張含有 字型, 印刷術, 文字, 筆跡 的圖片

  自動產生的描述
* TF-IDF是一種用於資訊檢索與文字探勘的加權技術，是一種統計方法，目的是想要過濾在檔案中常見的詞語，保留重要的詞語。
* TF-IDF是由包含TF（term frequency）及IDF（inverse document frequency）兩個部分。

1. TF: 單詞出現在一份文件的頻率
2. IDF: 單詞對於語料庫的重要程度

* 如果一個詞在一個特定檔案中具有高詞頻，在所有的檔案集合中具有低檔案頻率，則會有高的TF-IDF權重：

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

*答覆：*

* 以V1欄位作為label，spam為垃圾訊息、ham為正常訊息。

我做出來的結果：Accuracy Ratio為94.53%

一張含有 文字, 字型, 行, 數字 的圖片

自動產生的描述

我是用分類樹Decision Tree來訓練模型並分類，樹狀圖呈現於下方。

一張含有 圖表, 行, 地圖, 方案 的圖片

自動產生的描述

〈略加說明〉

首先以「憯」作為關鍵字，若這個詞「憯」在這句話中的重要性（Ex. TF-IDF）比重占超過0.097，則有可能是垃圾信件。

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

檢視全部資料，有包含「憯」這個詞的語句共257筆，僅4筆為ham正常訊息（4/257=1.56%），其餘253筆為spam垃圾訊息（253/257=98.44%）。因此若有一筆新資料，當中包含「憯」這個詞，則歸類為垃圾訊息的機率很高。

其它重要的關鍵字包括：憯 ≤ 0.097、www ≤ 0.111、p ≤ 0.103、mins ≤ 0.184、claim ≤ 0.135、nokia ≤ 0.154、service ≤ 0.504、uk ≤ 0.502、reply ≤ 0.796，若通過了以上檢測，才很有可能是ham正常訊息。

一張含有 文字, 螢幕擷取畫面, 字型, 行 的圖片

自動產生的描述

※ 已將process流程儲存為111C71008-1.rmp

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%)。

〔簡略提一下前處理，方便閱讀〕

* 刪去Row ID、Clothing ID及Review Text為Missing的欄位
* 設定Rating為Label，預測評分  
  Rating: Positive Ordinal Integer variable for the **product score granted by the customer from 1 Worst, to 5 Best**.