

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!

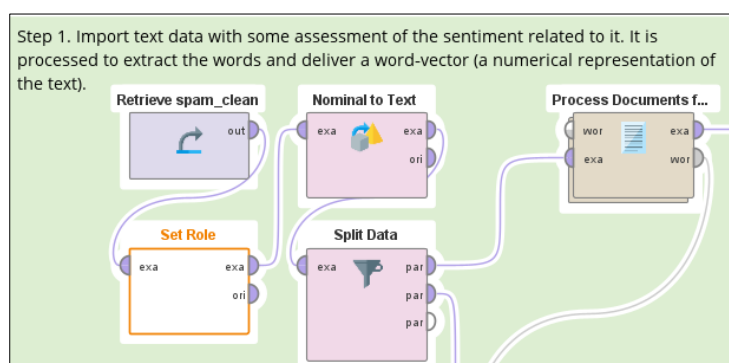
為了確保大家的學習品質,因此課程中一半的內容將做為作業的一部份;另一部份

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%)

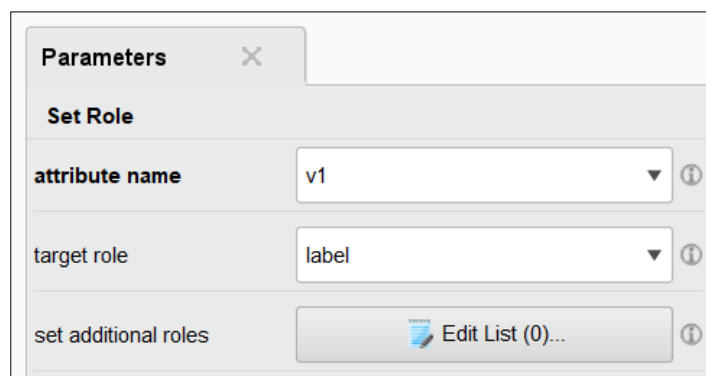
〔前處理〕



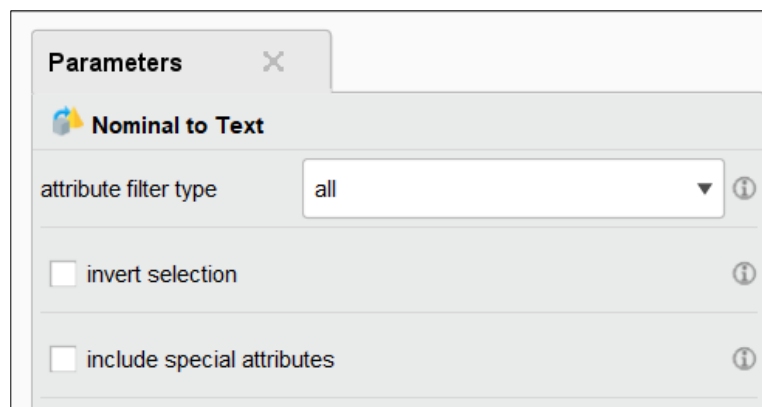
- A. Retrieve spam_clean : 讀取 spam_clean.csv 資料

Data 放置於資料夾內 · 檔名為 spam_clean.rm hdf5table

- B. Set Role : 設定 v1 欄位為 label



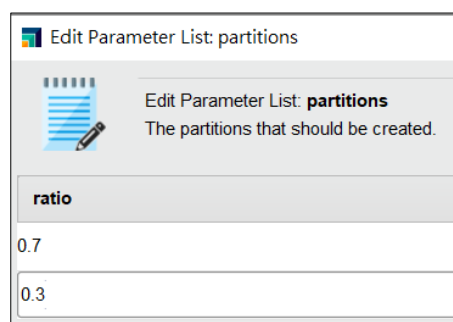
C. Nominal to Text : 資料轉為 text 型式



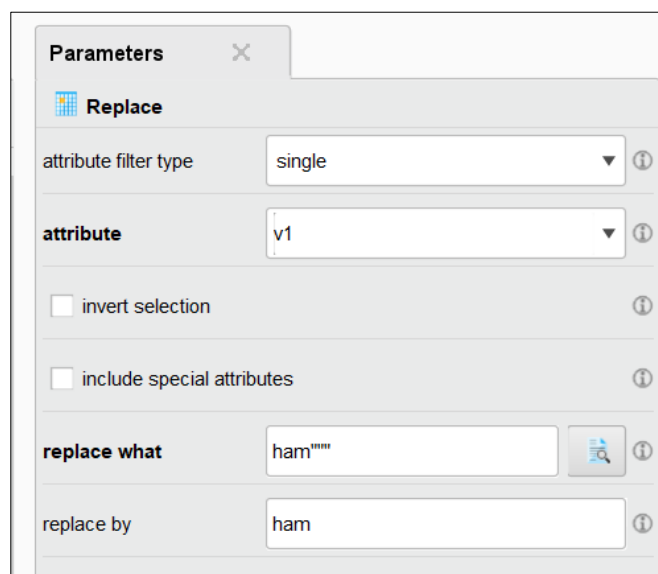
※ 由於讀入的資料其 v2 欄位屬性為 object，因此要轉 text

※ 使用 Nominal to Text 將 v2 轉換為 text

D. Split Data : 切分資料，訓練:驗證=70%:30%



E. (後續補充) Replace : 其中有兩筆資料讀成 ham''''，改為 ham。

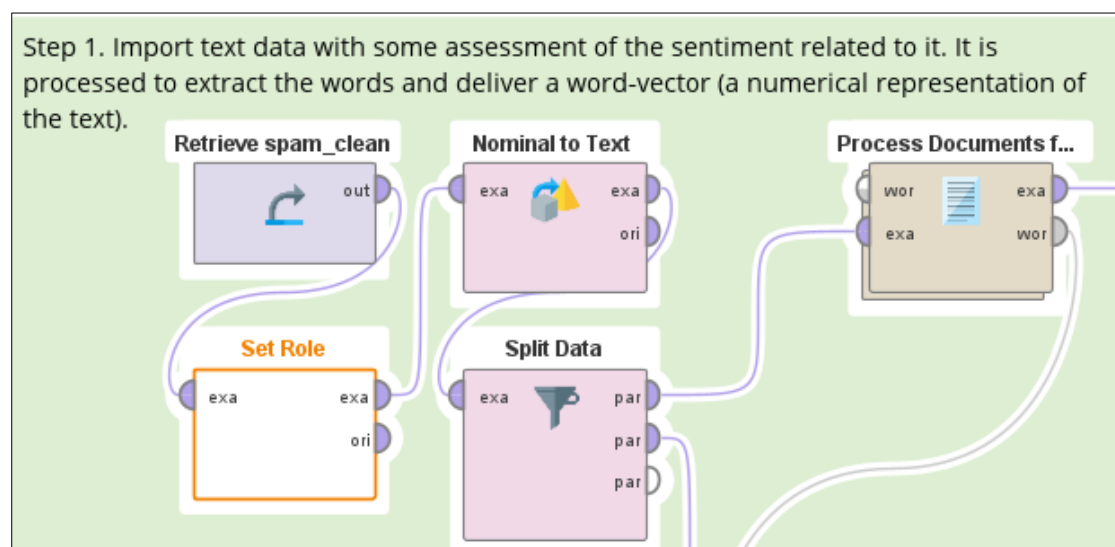


2. 接下來,我們要針對文字進行分析,因此有一些必要的處理程序,請說明一下您的步驟(10%),並請說明下列分析結果的意涵: TF-IDF (5%)

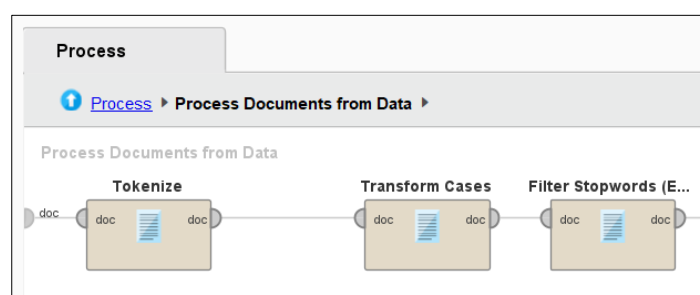
1. Step 1.

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。



※ 使用 Process Documents from Data 來進行文字的處理



- 使用 Tokenize 進行斷詞斷句，亦可解釋為將文本拆分成單詞。

- 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。

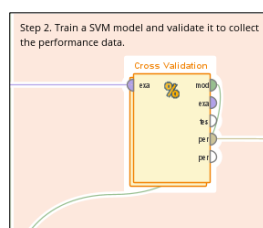
※（延伸解釋）什麼是停用詞（Stop Words）：

指一些頻繁出現、無具體含義或無助於文本分析的詞語，這些字詞對於文本並不具有明確的貢獻。在中文中常見的停用詞包括「的」、「和」、「在」、「是」、「了」、「及」、「或」等。

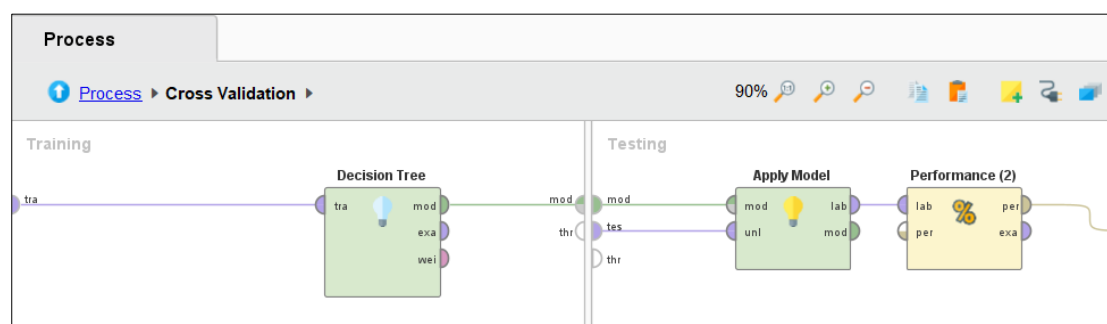
2. Step 2.

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

訓練模型



←右圖 · Cross Validation (交叉驗證)

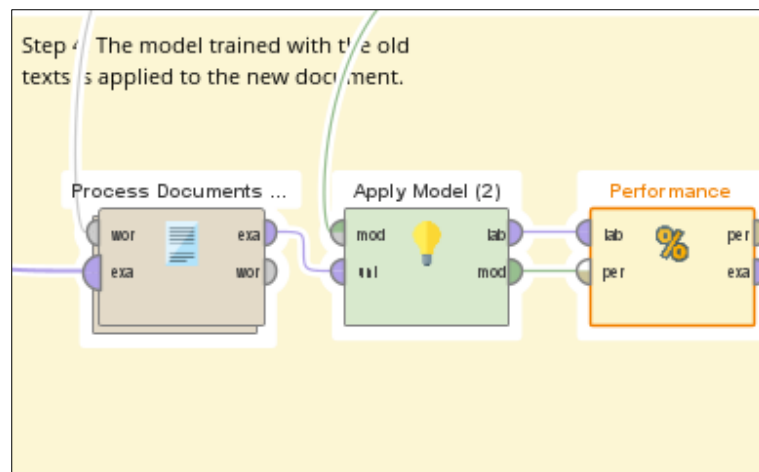


↑上圖，這裡以 Decision Tree 為例。

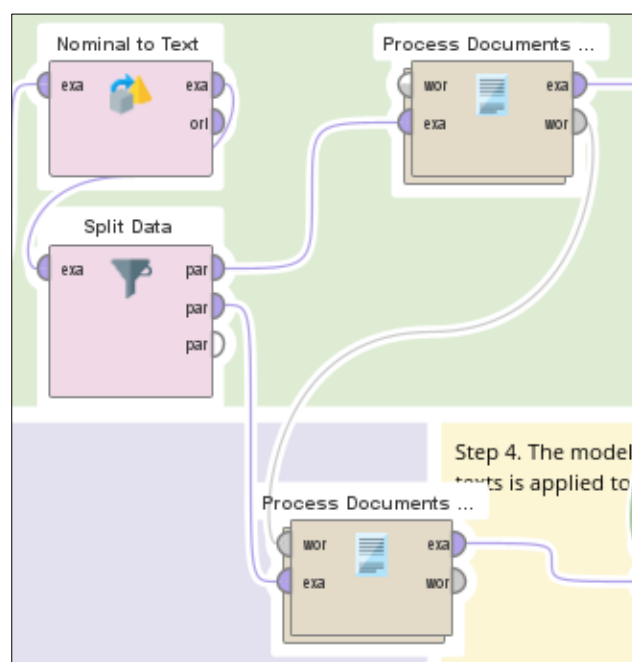
3. Step 3.

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

用訓練模型應用於 Testing Data。

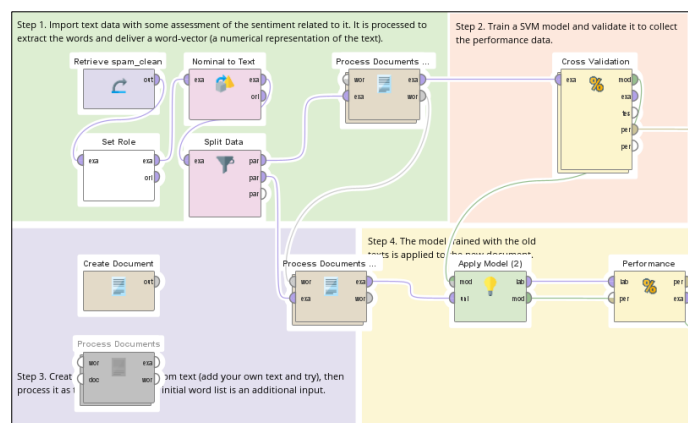


- Testing Data 也要做斷詞斷句處理 (Process Documents to Data) ，並且要給 word ，讓它知道如何斷句。
- 使用 Apply Model 將訓練完的模型拿去與 Testing Data 進行測試
- 接上 Performance 來看模型的結果，衡量正確率。



↑ 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



$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

✓ TF-IDF

- 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。

A. TF: 單詞出現在一份文件的頻率

例如：十萬 / 青年 / 十萬 / 肝，「十萬」出現了 2 次，總共有 4 個字

詞，因此 $TF = 2/4 = 0.5$

B. IDF: 單詞對於語料庫的重要程度

總共有 10 個文件，其中有 5 個文件中有提及「十萬」這個字詞，因此

$IDF = \log 10/5 \approx 0.3$

- 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

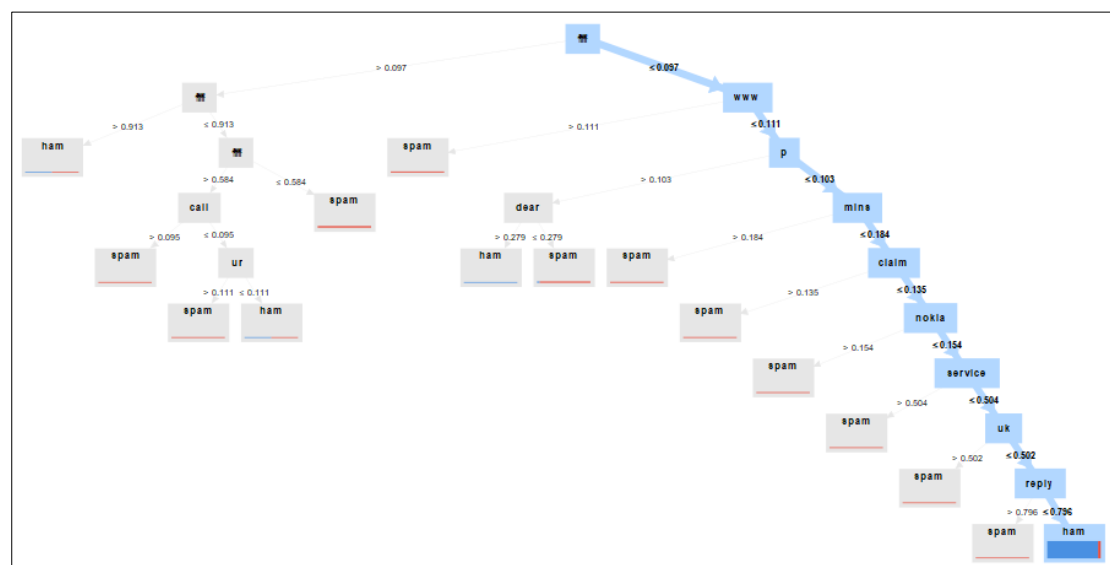
答覆：

- 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%

accuracy: 94.53%			
	true spam	true ham	class precision
pred. spam	135	8	94.41%
pred. ham	62	1075	94.55%
class recall	68.53%	99.26%	

我是用分類樹 Decision Tree 來訓練模型並分類，樹狀圖呈現於下方。



〈略加說明〉

首先以「慊」作為關鍵字，若這個詞「慊」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。

111C71008 創新 AI 碩一 何哲平

	A	B
1	v1	v2
7	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send 僣1.50 to rcv
10	spam	WINNER!! As a valued network customer you have been selected to receive a 僣900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
14	spam	URGENT! You have won a 1 week FREE membership in our 僣100 000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7R
36	spam	Thanks for your subscription to Ringtones UK your mobile will be charged 僣5/month Please confirm by replying YES or NO. If you reply NO you will not be charged
67	spam	As a valued customer I am pleased to advise you that following recent review of your Mob No. you are awarded with a 僣1500 Bonus Prize call 09066364589
69	spam	Urgent UR awarded a complimentary trip to EuroDisinc Trav Aco&Entry41 Or 僣1000. To claim txt DIS to 87121 18+6*僣1.50(moreFrmMob. ShrAcomOrSglSuplt)10 LS1 3AJ
95	spam	Please call our customer service representative on 0800 169 6031 between 10am-9pm as you have WON a guaranteed 僣1000 cash or 僣5000 prize!
115	spam	GENT! We are trying to contact you. Last weekends draw shows that you won a 僣1000 prize GUARANTEED. Call 09064012160. Claim Code K52. Valid 12hrs only. 150ppm
118	spam	You are a winner U have been specially selected 2 receive 僣1000 or a 4* holiday (flights inc) speak to a live operator 2 claim 0871277810910p/min (18+)
122	spam	URGENT! Your Mobile No. was awarded 僣2000 Bonus Caller Prize on 5/9/03 This is our final try to contact U! Call from Landline 09064019788 BOX42WR29C 150PPM
135	spam	Sunshine Quiz Wkly Q! Win a top Sony DVD player if u know which country the Algarve is in? Txt ansr to 82277. 僣1.50 SP:Tyrone
161	spam	You are a winner U have been specially selected 2 receive 僣1000 cash or a 4* holiday (flights inc) speak to a live operator 2 claim 0871277810810
168	spam	URGENT! We are trying to contact you. Last weekends draw shows that you have won a 僣900 prize GUARANTEED. Call 09061701999. Claim code S89. Valid 12hrs only
189	spam	Please call our customer service representative on FREEPHONE 0808 145 4742 between 9am-11pm as you have WON a guaranteed 僣1000 cash or 僣5000 prize!
226	spam	500 New Mobiles from 2004 MUST GO! Txt: NOKIA to No: 89545 & collect yours today! From ONLY 僣1 www.4-4c.biz 2optout 087187262701.50gbp/mtmsg18
313	spam	Think ur smart ? Win 僣200 this week in our weekly quiz text PLAY to 85222 now! T&Cs WinnersClub PO BOX 84 M26 3UZ. 16+. GBP1.50/week
336	spam	Valentines Day Special! Win over 僣1000 in our quiz and take your partner on the trip of a lifetime! Send GO to 83600 now. 150p/msg rcvd. CustCare:08718720201.
402	spam	FREE RINGTONE text FIRST to 87131 for a poly or text GET to 87131 for a true tone! Help? 0845 2814032 16 after 1st free tones are 3x僣150pw to e僣nd txt stop
419	spam	FREE entry into our 僣250 weekly competition just text the word WIN to 80096 NOW. 18 T&C www.txttowin.co.uk
425	spam	URGENT! Your Mobile number has been awarded with a 僣2000 prize GUARANTEED. Call 09058094455 from land line. Claim 3030. Valid 12hrs only
456	spam	Loan for any purpose 僣500 - 僣75 000. Homeowners + Tenants welcome. Have you been previously refused? We can still help. Call Free 0800 1956669 or text back 'help'
506	spam	+123 Congratulations - in this week's competition draw u have won the 僣1450 prize to claim just call 09050002311 b4280703. T&Cs/stop SMS 08718727868. Over 18 only 150ppm
516	spam	You are guaranteed the latest Nokia Phone a 40GB iPod MP3 player or a 僣500 prize! Txt word: COLLECT to No: 83355! IBHld Ldn W15H 150p/Mtmsgrcvd18+

檢視全部資料，有包含「僣」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「僣」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：僣 ≤ 0.097 、www ≤ 0.111 、p ≤ 0.103 、mins \leq

0.184、claim ≤ 0.135 、nokia ≤ 0.154 、service ≤ 0.504 、uk ≤ 0.502 、reply \leq

0.796，若通過了以上檢測，才很有可能是 ham 正常訊息。

ham
<div> <div>ham</div> <div> Distribution: 2520 <i>ham</i>, 127 <i>spam</i> </div> </div> <div> Number of items: 2647 </div> <div> Ratio of total: 88.62% </div>

※ 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藉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%)。

答覆：

[想法]

藉由分析評論 (Review Text 欄位) 來判斷顧客的意見是正面或負面

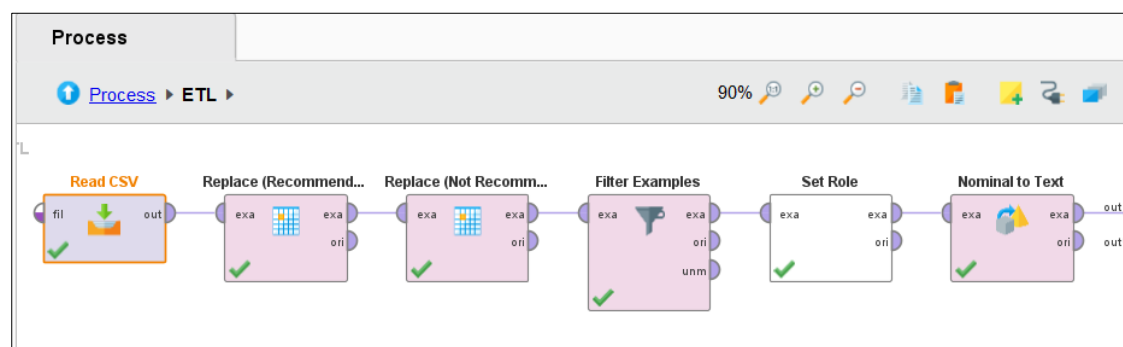
(Recommended 推薦此項產品→正面 ; Not Recommended 不推薦此項產品→負面)

※ Recommended IND 欄位的意思：

表示客戶是否推薦此項產品，1 為推薦，0 為不推薦。

(Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.)

[簡略提一下前處理，方便閱讀]



※ 模仿 spam_clean.csv 資料格式

- 在讀取時，僅讀取 Review Text 及 Recommended IND 欄位

column index	attribute meta data information
0	att1 <input type="checkbox"/> column s... polynomi... attribute
1	Clothing ID <input type="checkbox"/> column s... polynomi... attribute
2	Age <input type="checkbox"/> column s... integer attribute
3	Title <input type="checkbox"/> column s... polynomi... attribute
4	Review Text <input checked="" type="checkbox"/> column s... polynomi... attribute
5	Rating <input type="checkbox"/> column s... integer attribute
6	Recommendation <input checked="" type="checkbox"/> column s... polynomi... attribute
7	Positive Feedback <input type="checkbox"/> column s... integer attribute
8	Division Name <input type="checkbox"/> column s... polynomi... attribute
9	Department Name <input type="checkbox"/> column s... polynomi... attribute
10	Class Name <input type="checkbox"/> column s... polynomi... attribute

- 將 Recommended IND 欄位中的 1 取代為 Recommended; 0 取代為 Not

Recommended。

- 刪 Review Text 為 Missing 及 Recommended IND 不包含

Recommended。

- 設定 Recommended IND 為 Label

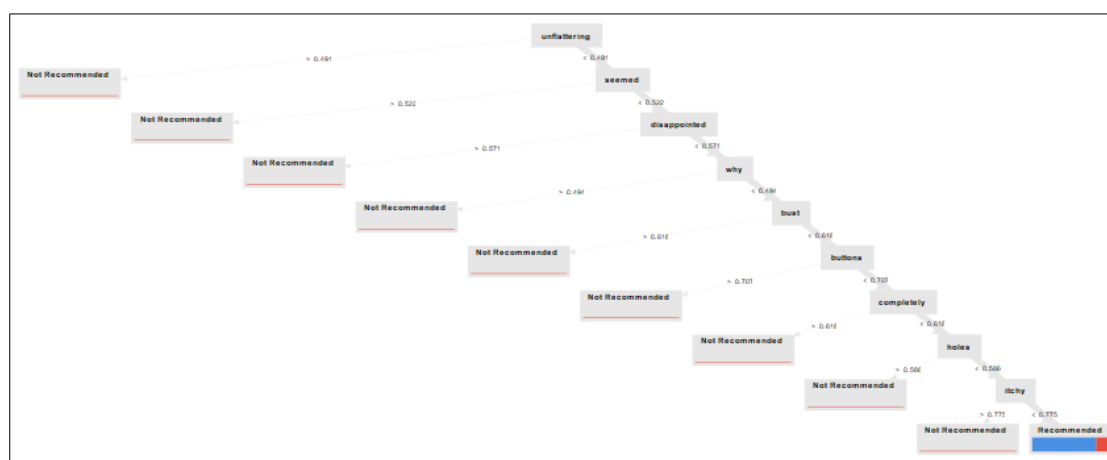
- Nominal to Text : 資料轉為 text 型式

〔說明分析結果〕

- 我做出來的結果：Accuracy Ratio 為 81.83%

accuracy: 81.83%			
	true Recommended	true Not Recommended	class precision
pred. Recommended	4973	1099	81.90%
pred. Not Recommended	8	12	60.00%
class recall	99.84%	1.08%	

- 我是用分類樹 Decision Tree 來訓練模型並分類，樹狀圖呈現於下方。



- 首先以「unflattering」作為關鍵字，若這個詞「unflattering」在這句話中的重要性 (Ex. TF-IDF) 比重占超過 0.491，則有可能是負面評論。

Recommended IND	Review Text
Not Recommended	Runs big and looked unflattering. i am petite, might work on someone taller.
Not Recommended	First, the fabric is beautiful and lovely for spring and summer. i really wanted to like this top, but the fitting is so awkward for me. i typically wear a 0's, and sized up in the shirt to a size 2. it was very tight and pulled funny.
Recommended	I am obsessed with peplum down coats because the ones you usually see have no shape and are extremely unflattering. i was excited for this to arrive. this is quite nice and it looks more feminine than the other down coats out.
Not Recommended	True to size on the neckline and arms but extremely large and puffy in the torso. very unflattering cut!
Not Recommended	This top is good quality and cute. it runs large- i'm usually a medium and needed a small. the reason i will be returning it is because it flares out at the bottom on the black which is very unflattering on. it makes me look wide.
Recommended	I got the green color with gray accent stitching, looks awesome with gray tone leggings. super soft, definitely runs a little big (5'6" 140 lbs i ordered the small) but not in an unflattering way.
Recommended	Based on some reviews i decided to get the regular xs, even tho i am an xs petite (5'2, 107 lb, 32c) i found the fit to be flattering -- fitted enough but not too loose or tight. the length is perfect and work appropriate, and the mat.
Not Recommended	It looks like you are wearing cargo shorts. really unflattering. avoid buying this skirt.
Not Recommended	Loved the style, ordered my normal m-fit but i thought i would potentially taper the sides to make less boxy (with a larger bust and shoulders, boxiness is unflattering). the slight boxiness looked perfect with stretchy, fit.
Not Recommended	I am floored by the amount of positive reviews on this dress! when i received it, it looked nothing like it does on the model. the bottom looked like dirty sand and was completely wrinkled. if you have anything above a c cup,
Not Recommended	I ordered this sweater in black and thought it was a pretty design and something different in a cardigan. when i tried it on, it was an odd fit and very boxy and unflattering. i wanted to like it and keep it, but i knew i would not.
Recommended	I am 5 feet and 120. i ordered a petite small in moss. i have posted a photo below. it's a pretty green. the lining is pretty, the reason i deducted a star is because the lining is very thin and 'slips' a little when putting the jacket on.
Recommended	Very comfortable dress, with a gorgeous pattern. it comes with a self liner in navy that is very smoothing - i was worried a figure hugging sweater knit would be unflattering. cut is perfect. i'm a size 10/12 on top and 12/14 on
Not Recommended	This dress was pretty but had a weird fit. the waist droops down in the middle instead of going straight across, which i found unflattering. i ended up returning it.
Not Recommended	This was so unflattering. the bust was too high so it hit me weird on the boobs. the green was a beautiful, vibrant color but the fabric is so delicate that one wrong move would cause a snag.
Not Recommended	I had reservations about this top based on other reviews, but it looked cute on the model so i took my chances. i have an athletic build-not model-thin, but not dumpe either-and this top made me look very frumpy and matron
Not Recommended	I loved this top in the blue and wanted so badly for it to fit however it was very unflattering on my 5'7 frame. the bottom hemline is straight across all around whereas i thought it was longer in the back. i could see this looking
Not Recommended	I wanted to love this dress, and thought it would be perfect for a barn wedding i have coming up. it was allllllmost right but sadly fell short. the bib that hangs over the bust hits in a really unflattering mid point that makes you
Not Recommended	I purchased this shirt in grey. i loved the color and the details, especially the lace on the shoulders. i just found the fit to be off for me. the style certainly calls for a relaxed fit but the small was loose in an unflattering way. had
Not Recommended	Although i love the soft feel of the sweater, the zig zag design was very unflattering. also, the length is much shorter than appears in the picture. i will be returning it.
Recommended	Very cute dress but the skirt flares out more than it looks in the picture. this made my midsection look unflattering from the side. also, for some who is on the flat chested side, you need a good bra or alteration (at least for me)
Recommended	This dress had so much promise from the picture. i loved the midi length, color, and fascinating waist detail. i ordered online, so there was no way to try on before purchasing. the fit was disappointing. it fell in an unflattering
Recommended	I love this top. it is very short and hits at a potentially unflattering area if one has wider hips. the fabric and swing to the 'skirt' is adorable.

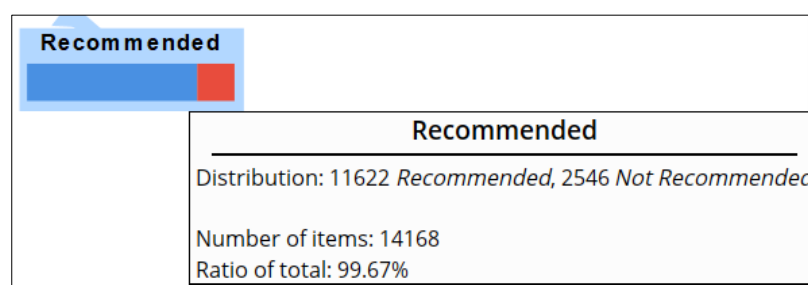
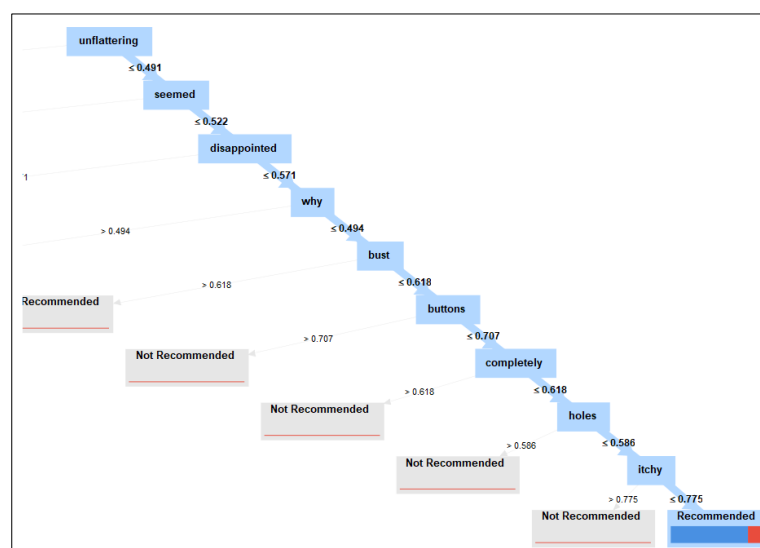
檢視全部資料，有包含「unflattering」這個詞的語句共 299 筆，其中

209 筆為負面評論 (209/299 約為 70%)，剩餘 90 筆為正面評論

($90/299=30\%$) 。因此若有一筆新的顧客評論，當中包含

「unflattering」這個詞，則歸類為負面評論的機率較高。

- 其它重要的關鍵字包括： $\text{unflattering} \leq 0.491$ 、 $\text{seemed} \leq 0.522$ 、 $\text{disappointed} \leq 0.571$ 、 $\text{why} \leq 0.494$ 、 $\text{bust} \leq 0.618$ 、 $\text{buttons} \leq 0.707$ 、 $\text{completely} \leq 0.618$ 、 $\text{holes} \leq 0.586$ 、 $\text{itchy} \leq 0.775$ ，若通過了以上檢測，才很有可能是正面評論。



- ✓ 另外可以再深入分析，探討包含年齡、產品部門與正面評論的彼此關係。

例如哪一年齡階層給予最多正面評論，或是哪一個產品部門獲得最多的正面評論。

- ※ 已將 process 流程儲存為 111C71008-2.rmp