Information Retrieval and Extraction Term Project2
Relation extraction

一、Team member：

湯忠憲 資料科學碩一 R06946003

劉宏國 資工碩一 R06922006

陳奎伯 資工碩一 P06922001

Agree to share your report with your classmates? (YES)

二、Methodology：

為提取訓練用的資料，我們使用列的演算法從已經分好詞的紅樓夢文本中提取相關的字句：

Sentences Extraction Algorithm

Algorithm1

Input: Entity pairs and Corpus
for each entity pair:
  for each sentence in Corpus:
    if two entities in same sentence: (Condition 1)
      extract this sentence
      break
    else if two entities in context (appear together with 3 sentences): (Condition 2)
      extract context
      break
    else: (Condition 3)
      extract two sentences where they first appear in corpus, separately
      break

從 Algorithm1 中可以看到提取的句子有三種來源，並且 Condition 1 下提取的句子品質最好，Condition 3 的最差。各 condition 於 training set 的個數分別為 66、48 和 36。可以看到會有許多品質不好的訓練資料。

Classification Methodology and Experiments

方法一：Random Forest

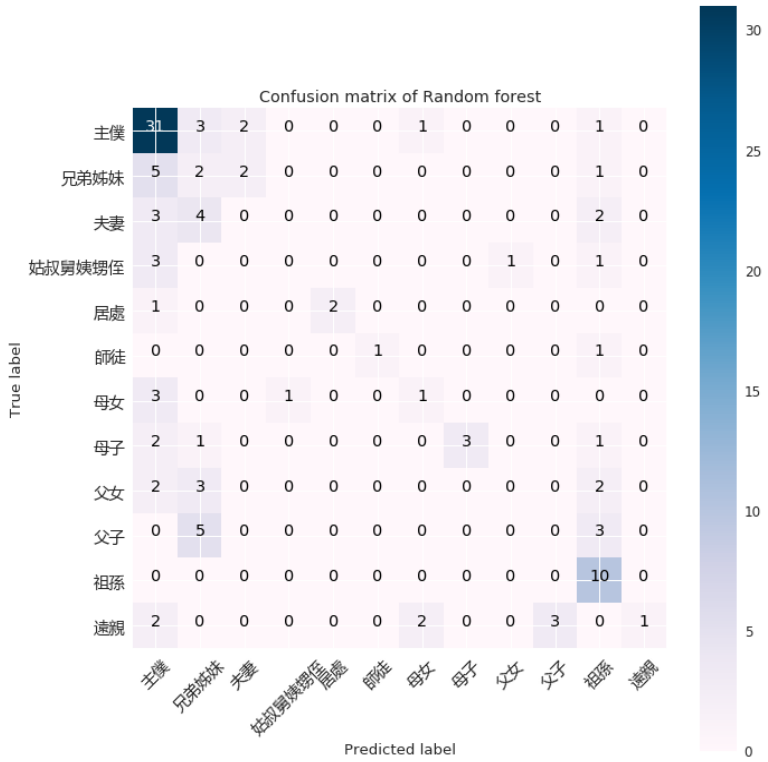
首先我們將 relation extraction 的問題考慮成一種分類問題，也就是將兩個 entities 的關係分成 12 個類別。從 Algorithm1 中我們能將 entities 出現的句子提取出來做為訓練資料，並且利用 Word2Vec 賦予每個詞一個向量，再將單一句子中的這些向量平均起來成為用來表示該句子的向量。這邊我們將句子投射成 100 維的 representation vector。

接著，有了數值化的訓練資料後，我們利用隨機森林(random forest)作為分類器，將每筆訓練資料分為 12 種關係。

分類器參數：*n_estimators=100, max_depth=8, max_features='sqrt'*

準確度約為：0.46428

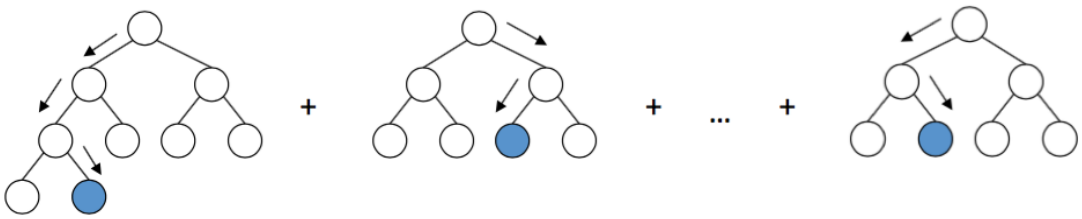
圖一為使用 random forest 作為分類器所產生的 confusion matrix。可以看到**主僕**關係的準確率還蠻高的，因為 training data 中有不少**主僕**關係。相對的，因為 training 資料的 highly-unbalance 蠻多其他關係也會被分到主僕關係的。次好的類別是祖孫關係。



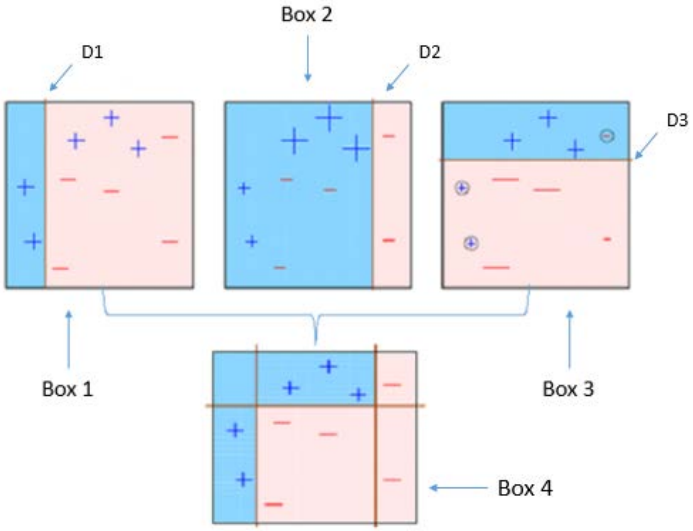
圖一、Confusion matrix of Random forest

方法二：Gradient Boosting Tree (implementation: xgboost)

我們還有嘗試使用 Gradient Boosting Tree 的一種 implementation，XGBoost。Kaggle 上蠻多人用的。該算法類似 Random Forest，即建構很多棵決策樹來提升準確度以及穩定性(如圖二)，不一樣的是他透過 sequential learning 的方式，補足上一次分類的錯誤（如圖三，Box 2 修正 Box 1，Box 3 修正 Box 2，以此類推)。同時由於該模型實現樹的 Regularization，即使樹的深度加大，它也能夠一定程度的防止 overfitting。



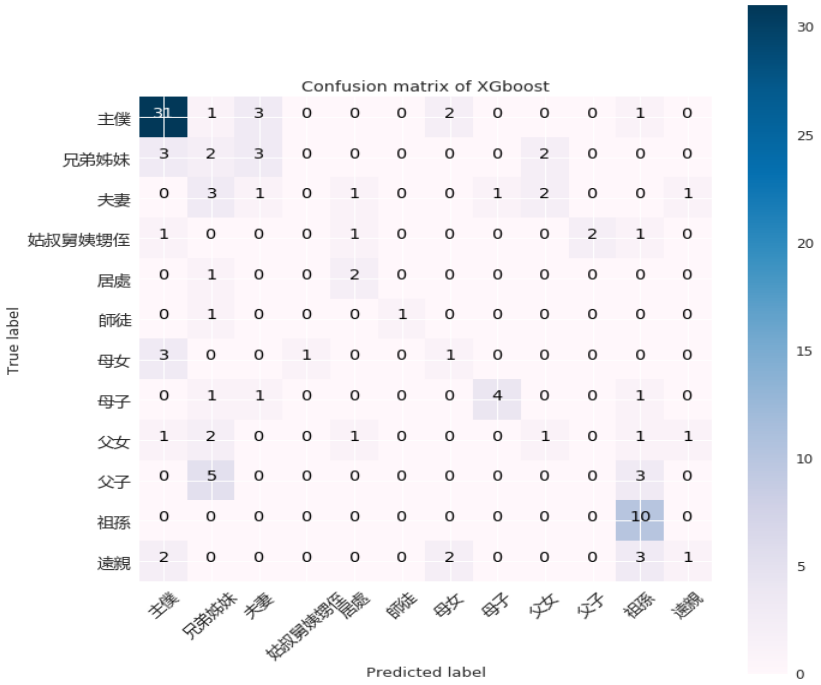
圖二、Ensembling of Decision Tree



圖三、Boosting mechanism

分類器參數：*max_depth=7, n_estimators=10*

準確度約為：0.48212



圖四、Confusion matrix of XGBoost

方法三：Rule based

這個方法我們使用 handcraft 的 feature 來分類關係，將任務考慮成一種 pattern extraction problem。以下是實作的步驟：

Step1：建立 rule based

一開始會先算 sentence 中，每一個字的 tf-idf 值，並取出 weight 大於 0.01 的字。再將那些字扣掉 stopwords 並且對應 training data 中的 label 建立 rules。因此 rule based 的形式大致來說就如下：

relation_word[2][0:10]：(index=2 代表夫妻)，夫妻相關 term。
['見', '獨', '堪', '倒', '身', '恐', '忙', '十分', '做', '賈珍']

relation_word_weight[2][0:10]：(index=2 代表夫妻)，夫妻相關 term 的 weight。
[0.04, 0.13, 0.17, 0.07, 0.09, 0.38, 0.15, 0.13, 0.10, 0.11]

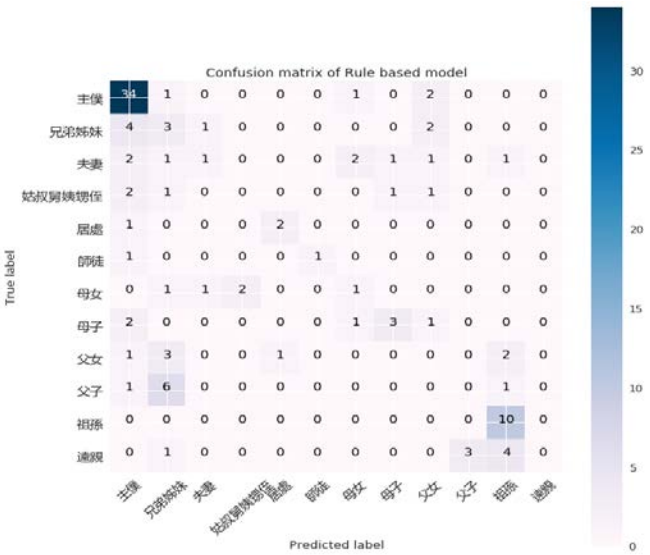
像是**見**這個 term 在 rule based 中是屬於夫妻這個關係，且它的 weight 是 0.04。

Step2：

抽取出 testing 的句子一樣會先去算每一個字的 tf-idf 值，也取出 weight 大於 0.01 的字並扣掉 stopwords，那因為一個句子所抽取出來的 term，有些在 rule based 中可能是在**夫妻**這個 label 之中，另一些可能是在**主僕**之中，因此會分別算 12 種 **relation** 得到的分數。例如：一句 **sentence** 中假如抽取出來的 term 包括：**膝、皮**在 rule based 中屬於**夫妻**這個 relation；**命令**在 rule based 中屬於**主僕**這個 relation，因此夫妻這個 relation 的分數為：**膝**的 term weight + **皮**的 term weight；而**主僕**得到的分數為：**命令**的 term weight。最後在看哪個 relation 的分數最高，則那兩個人名 entity 就為該 relation。

模型參數：取出 weight 大於 0.01 的字當作比對 pattern（關鍵字）

準確度約為：0.49107

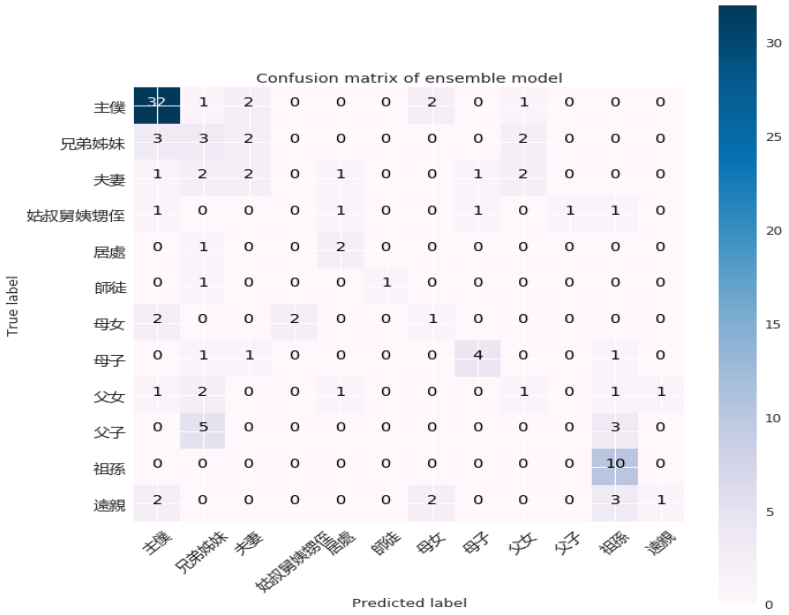


圖五、Confusion matrix of Rule based model

方法四：Rule based + XGBoost

此方式是將 Rule based 和 XGBoost 做結合，透過 XGBoost 預測 relation 的機率取最高 relation 的機率，在看該最高的機率是否大於 0.3，如果大於 0.3 則該 relation 就使用 XGBoost 預測出來的結果；小於等於 0.3 則使用 Rule based 方法預測出來的結果。

準確度約為：0.50892



圖五、Confusion matrix of ensemble model

方法五：Rule based + Entity Class

此方式是從紅樓夢原始文章及訓練資料擷取人物的相關資訊如性別、父親、母親、兄弟姊妹、祖孫、出現次數頻率等。每個人物的型態是以 Entity Class 做為表示，上述各項資訊則記錄於 Class 的對應 Attribute。這個方式的優點是可以透過推導的方式得知人物間接的關係，例如 A 與 B 是兄弟姊妹，且 B 與 C 是兄弟姊妹，則 A 與 C 也是兄弟姊妹。另外根據紅樓夢當代世界觀，定義了一些重要字元集來增加判斷資訊：如四大家族姓氏為賈史王薛、婢女常稱作丫頭等。

實驗編號	移除項目	準確度
1	Entity Class 屬性資料	0.089286
2	四大家族姓氏字元集	0.580357
3	地點相關字元集	0.598214
4	主僕相關字元集	0.339286
5	無	0.625000

實驗得知「Entity Class 屬性資料」以及「主僕相關字元集」是最重要的 feature

準確度為：0.6250

三、Discussion：

Random Forest	0.46428
XGBoost	0.48212
Rule based	0.49107
Rule based + XGBoost	0.50892
Rule based + Entity Class	0.62500

表一、各方法及其準確率

從上述的模型準確率(表一)和 confusion matrix 的比較中可以發現，各 label 的準確度與訓練資料的數目有很大的關係。其中，**主僕**關係的精準分類為準確度的主要貢獻。若是有更多資料的話，可以考慮使用 LSTM 直接訓練一組 sentences embedding，可能準確率會比較高。此外，Rule based (pattern extraction) 的方法在這邊能提供最好的準確率，該方法能有效利用 training data 中的 pattern 來分類。例如: **臊皮**表示輕薄、占便宜的意思，在紅樓夢中有許多丈夫心疼娘子被人臊皮的場景（見《紅樓夢·第二五回》），因此使用這個詞能輕易的將 entity pair 的關係定為**夫妻**，這可說是結合中國文學以及電腦科學於本年度重大的發現。但這樣的方法容易因為詞彙多變而降低準確度，若是再加上擷取保存人物的相關資訊如性別、父親、母親、兄弟姊妹、祖孫、出現次數頻率等，並依照當代世界觀預先定義判別關係的詞彙如**丫頭**，則準確率可以再進一步提升。