Model III. 負責人: 王嚴徵

**實作語言與套件:**

Python3(64-bit), scikit-learn, numpy

**實作過程與結果討論:**

一開始先將兩個人物進行特徵抽取，首先需要擷取每一種關係的用語特徵，使用處理好斷詞以及標註好詞性的文本，以30句話為window(測試的結果發現會較7, 20, 40句為單位作為window好)，只要兩個人同時在這個範圍內出現，便將這30個句子存放到對應的關係當中，因此可以分別對12種關係抽取出許多句子，並利用TFIDF值排序找出這12種關係的前100個關鍵字詞，作為用語模式的特徵(Ex:有上對下關係的配對，像是父母與子女、主僕，句子中便可能會出現像是「進來」、「丫頭」、「過來」等命令式或上對下的稱呼詞)，這邊有特別將紅樓夢中常出現的人物「賈寶玉、王熙鳳、林黛玉、薛寶釵、史湘雲、賈探春、賈迎春、賈惜春、賈元春」，因為彼此之間常常僅以名互相稱呼，所以也會去掉姓氏進行查找。

接下來針對訓練資料與測試資料，一樣以30句為window取出所有共同出現的句子，利用之前取出的所有關鍵字以出現次數Count矩陣建立特徵(以Count矩陣，準確率會較TFIDF矩陣來的高)。之後使用Random Forest(RF)演算法進行分類(有試過SVC以及GBDT，但調過參數都蠻容易overfitting，所以之後都只考慮RF)，可以得到33%的準確度(主僕這個類別大多可以答對，因為訓練資料多，且容易將其他有上下關係的類別也猜成主僕)。

因為紅樓夢人物眾多，許多人物之間可能並沒有共同出現的句子，所以另外使用姓氏作為特徵，以訓練與測試及資料的所有人物，找出四大家賈史王薛之外的常出現姓氏「李、林、金、秦、邢、尤」，加上剩下姓氏(或是無姓氏，像是僕人)以及居處(透過Nc篩選)共12種使用one-hot encoding作為特徵，可以將準確率提升到43.75%。稍微調整一下RF的參數，將max\_depth改為7, n\_estimator改為10，便可以做到50%的準確率。

從結果可以發現主僕、祖孫、遠親類別答對率很高，而其他父母與子女、夫妻、親戚關係很低。所以利用維基百科的資訊，手動標記所有人物的性別(居處及部分遠親為無)，以長度為3的特徵向量表示兩者性別關係，但是結果卻顯示訓練資料的準確率上升，但測試集資料準確率卻下降到47.3%，顯示性別並不是一個好的特徵。也有使用一個120長度的向量去表示兩個角色各自第一次出現的章節，但效果也不是很好。因為有許多人之間並無共同出現的段落，所以也會尋找每一個人第一次出現的句子，並抽取上下兩個句子一起作為用語特徵，讓沒有共同出現的配對也可以有用語的特徵，如此可以些微提升準確率到50.8%，之後再加入中文停用詞可以提升到51.8%。因為文本本身有標註詞性，所以最後再訂一個規則，如果兩個entity有其一個POS tag為”Nc”的話，直接將其分類到「居處」這一類，最終可以將準確率提升到54.46%。

|  |  |  |
| --- | --- | --- |
| Feature & Algorithm parameters | Ein | Eout |
| 用語特徵 +  RF (max\_depth=5, n\_estimator=5) | 56 % | *33 %* |
| 用語特徵 + SVC / GBDT | > 70% | *42%* |
| 用語特徵 + 姓氏特徵  RF (max\_depth=5, n\_estimator=5) | 65 % | *43.75 %* |
| 用語特徵 + 姓氏特徵  RF (max\_depth=7, n\_estimator=10) | 73 % | *50 %* |
| 用語特徵 + 姓氏特徵 + 性別特徵  RF (max\_depth=7, n\_estimator=10) | 80.6 % | *40.1 %* |
| 用語特徵 + 姓氏特徵 + 章節特徵  RF (max\_depth=7, n\_estimator=10) | 75.3 % | *47.3 %* |
| 用語特徵 + 姓氏&性別&章節特徵  RF (max\_depth=7, n\_estimator=10) | 83.3 % | *41.9%* |
| 用語特徵(含第一次出現) + 姓氏特徵  RF (max\_depth=7, n\_estimator=10) | 79.3 % | *50.8 %* |
| 用語特徵(含第一次出現，去停用詞)  + 姓氏特徵  RF (max\_depth=7, n\_estimator=10) | 80.6 % | *51.78 %* |
| 用語特徵(含第一次出現，去停用詞)  + 姓氏特徵 + 居處filter  RF (max\_depth=7, n\_estimator=10) | 80.6 % | *54.46 %* |

由最後的結果可以發現最佳的訓練模型在主僕、祖孫、遠親、居處等部分可以有很好的預測結果(錯誤皆少於2個)，其他部分的正確率則較低，在父母與子女、姑叔舅姨甥侄的部分可能是因為其用語可能也有較多的上對下命令語氣，加上主僕的訓練資料占比非常高，這兩者的資料相對少，所以還蠻容易將這些關係預測成主僕(這些類別在training data的預測錯誤率也很高)，比較特別的是夫妻在訓練集資料的預測還不錯，但是在測試籍資料完全沒有正確，有可能是因為測試集資料的夫妻之間的共同出現段落較少的緣故。由實驗結果可以知道要特徵擷取對於模型準確率相當重要，可以大幅提升準確率(姓氏特徵)。用語的句子抓取與關鍵字擷取部分我覺得是可以改進的地方(有使用Mutual information的方式尋找關鍵字，但效果不好)，因為從生活中便可以發現到不同的關係之間的用語非常不同(上對下、平輩、性別、熟悉與陌生、年輕人宇中老年人…等等)，或是可以增加訓練集資料以及使用常出現的人物作為資料可能也是改善準確率的方法之一，因為大部分有較多資料與擷取到較多共同句子類別的預測結果都還不錯。