

模
型
特
色

Task 1

使用結巴將台詞斷詞後選collocation最高(且出現次數高於3次)的bigram當電影名稱, 若沒有collocation則使用詞頻及詞性組合出電影名稱

Task 2

加入了是否具有反差詞之特徵幫助判斷電影名稱的吸引力; 以及將SVR及迴歸模型預測出的電影票房做加總後排名

NLP Final

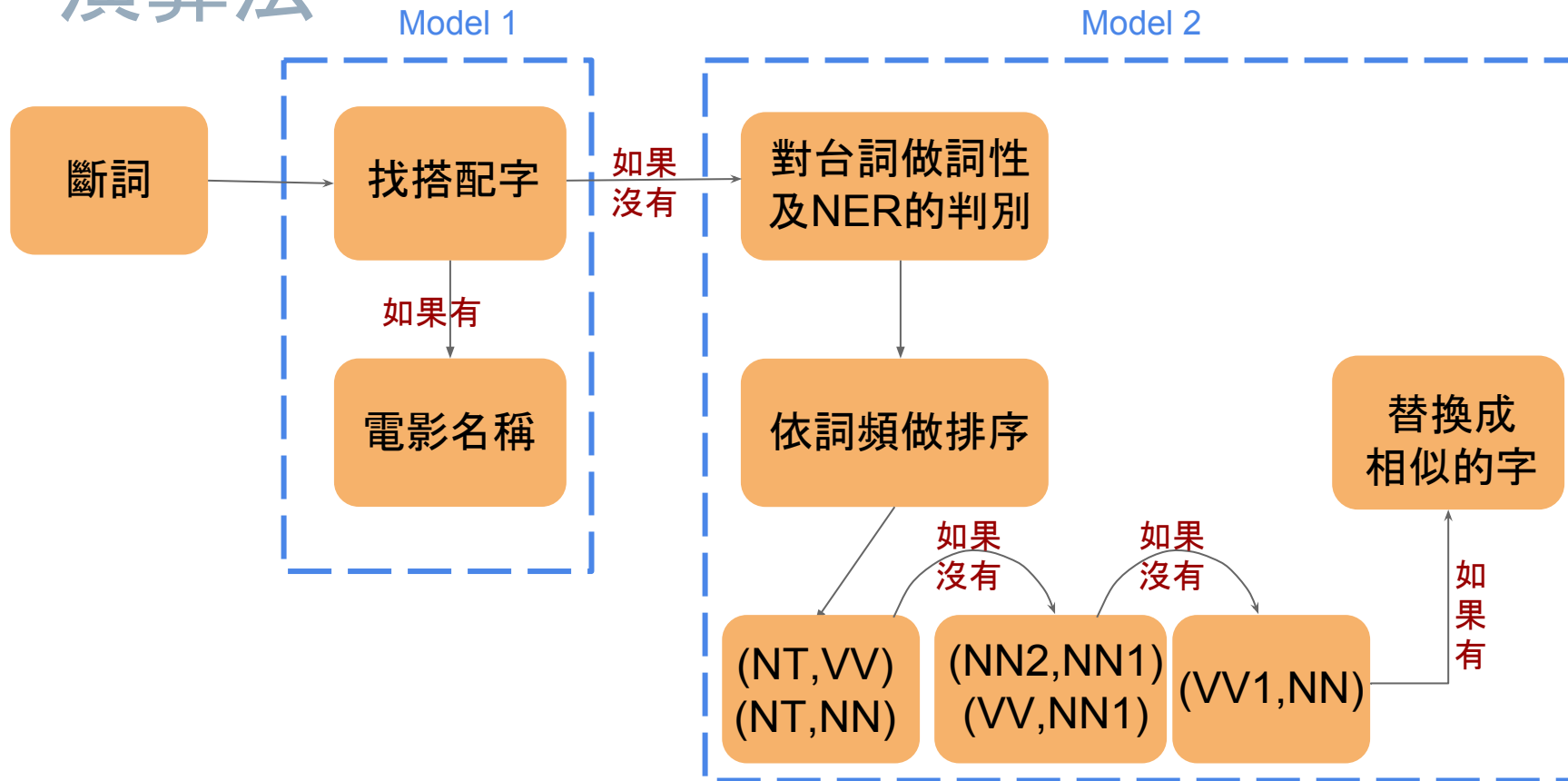
第5組

統計所/劉彥伶、莫惠淇、黃婉婷

Task 1 電影名稱生成



演算法



DEMO

原始電影名 || 生成電影名

神鬼奇航 世界的盡頭 || 幽冥飛船

控制 || 月日

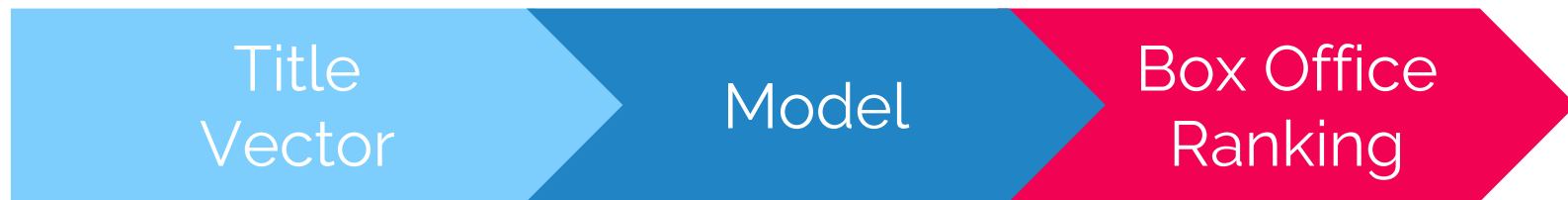
色戒 || 麥太太

唐山大地震 || 挺好

美女與野獸 || 克萊蒂

海角7號 || 馬拉桑

Task 2 為電影名稱打分數



Combine Embedding Vectors:


1. sum
2. average
3. concatenate

1. regression
2. SVR
3. random forest
4. neural network

Example — 冰雪奇緣 (sum of embedding vectors + regression)

- step1: 將「冰雪奇緣」斷詞成「冰雪」、「奇緣」
- step2: 找出「冰雪」、「奇緣」各自的k維詞向量 X_1, X_2, \dots, X_k
- step3: 將兩個詞的 X_1, X_2, \dots, X_k 各別相加, 相加後的結果即為「冰雪奇緣」的詞向量
- step4: 將「冰雪奇緣」的詞向量輸入模型得到預估票房

	X_1	X_2	\dots	X_k
冰雪	0.1	0.32	\dots	0.12
奇緣	0.01	0.25	\dots	0.3
冰雪奇緣	0.11	0.57	\dots	0.42
	X_{1i}	X_{2i}	\dots	X_{ki}


$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \dots + b_k X_{ki}$$

Task 2 最終版作法 (Linear regression + SVR)

Train Model

- 前置準備: 蒐集1012部電影名稱及其全球票房、多部電影台詞, 將所有蒐集到的文本用結巴斷詞, 使用word2vec找出每個詞對應的64維詞向量
- step1: 將電影標題轉為64維的向量(用加總的方式合併詞向量)
- step2: 計算電影標題由幾個字組成、含有幾種詞性的詞語、用結巴斷詞後的token數量、用coreNLP斷詞後的token數量、有無出現命名實體(有為1, 沒有為0)、有無出現**反差詞**(有為1, 沒有為0)
- step3: 使用電影標題向量訓練出一個可預測票房的linear regression模型
- step4: 使用step2計算的六種特徵訓練出一個可預測票房的SVR(linear kernel)模型

反差詞之定義:

將電影名稱斷詞後, 若有任一對詞其詞向量之 cosine similarity小於0, 則此電影名稱中有出現反差詞。
e.x.「美女與野獸」中,「美女」和「野獸」其詞向量之 similarity < 0, 故「美女與野獸」中有出現反差詞。
我們認為有出現反差詞的電影標題可能會特別吸引人, 所以在模型中加入此特徵。

Testing

- step1: 將電影標題轉為64維的向量(用加總的方式將詞向量合併, 如果找不到對應的詞向量用0取代)
- step2: 計算電影標題由幾個字組成、含有幾種詞性、用結巴斷詞後的token數量、用coreNLP斷詞後的token數量、有無出現命名實體(有為1, 沒有為0)、有無出現反差詞(有為1, 沒有為0)
- step3: 將電影標題向量輸入訓練好的linear regression模型、step2計算的六種特徵帶入訓練好的SVR模型
- step4: linear regression及SVR輸出的票房分別正規化後相加並排序

Q & A