自然語言處理 實作一

October 19, 2017

OUTLINE

- 注意事項
- Task 1
- Task 2
- Task 3
- Task 4
- 規則

注意事項

- 請先在自己的電腦跑測資，再把結果上傳至網路
- 處理過程中記得把Token都轉成小寫
- 請依照指定格式輸出測試結果
- 每題皆可重複繳交測試結果
- Demo結果請至以下連結查詢
- https://docs.google.com/spreadsheets/d/1HuNyHqg66CPqsgQrWrK3m1xfJI669l7W3U8urEjheFI/edit?usp=sharing

TASK 1

- 用一個dimension=300的vector來表示每個對話句
- 每一個dimension代表一個詞
- TF-IDF如下

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = frequency of i in movie line j
df_i = number of lines containing i
N = total number of lines

TASK 1 輸出要求

- 第一行為每個dimension所對應的單字
- 第二行為每個dimension的idf值
- 第三行之後為對話句的vector

```
dachshund | brown | dog | adorable | short | legs   (用" | "隔開)
1.4,0.3,0.2,0.5,0.7,0.6                              (idf值)
0,0,0,8,0,0                                          (對話的TF-IDF)
…
…
```

TASK 2

- 用一個dimension=300的vector來表示每個對話句
- 每一個dimension代表一個詞和詞性
- TF-IDF如下

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = frequency of i in movie line j
df_i = number of lines containing i
N = total number of lines

TASK 2 輸出要求

- 第一行為每個dimension所對應的單字
- 第二行為每個dimension所對應的詞性
- 第三行為每個dimension的idf值
- 第四行之後為對話句的vector

doge|such|wow|much|awesome	(用"|"隔開)
NN|JJ|UH|JJ|RB	(用"|"隔開)
1.4,0.3,0.2,0.5,0.7	(idf值)
1,0,0,8,0,0	(對話的TF-IDF)
…	
…	

TASK 3

- 分別用兩種計算Similarity的方式計算電影的相似度
- 每一種方法都得舉出三部相似的電影，以及一部不相似的電影
- 第一種請使用**Cosine Similarity** (參考3-a p.17)
- 第二種方法可以任意選擇(不限上課所學的)

TASK 3-1(COSINE SIMILARITY)

- 第一行為三部相似電影的名稱
- 第二行為電影dimension所對應的字
- 第三行到第五行為每部電影的vector
- 第六行為不相似電影的名稱
- 第七行為電影的vector

Thug Life | Feels Bad Man | The Legend 27 (用" | "隔開)
tough | gun | gang | frog | game | of | war (用" | "隔開)
…
…
…
Epic Sax Guy
. . .

TASK 3-2(自己的方法)

- 輸出格式比照cosine similarity的格式
- 請在上傳檔案的備註欄位說明用什麼方法
- 有疑問或不確定的可以隨時問TA

TASK 4

- 請先簡化每個句子、建立bigram model再計算Entropy；然後試著比較句子簡化前的bigram model的Entropy。

- 簡化過程如下頁所示

TASK 4(CONT'D)

原句：We will build a great wall along the southern border, and Mexico will pay for the wall.

Universal dependencies

```
nsubj(build-3, We-1)
aux(build-3, will-2)
root(ROOT-0, build-3)
det(wall-6, a-4)
amod(wall-6, great-5)
dobj(build-3, wall-6)
case(border-10, along-7)
det(border-10, the-8)
amod(border-10, southern-9)
nmod(wall-6, border-10)
cc(build-3, and-12)
nsubj(pay-15, Mexico-13)
aux(pay-15, will-14)
conj(build-3, pay-15)
case(wall-18, for-16)
det(wall-18, the-17)
nmod(pay-15, wall-18)
```

Universal dependencies

```
nsubj(build-3, We-1)
aux(build-3, will-2)
root(ROOT-0, build-3)
det(wall-6, a-4)
amod(wall-6, great-5)
dobj(build-3, wall-6)
case(border-10, along-7)
det(border-10, the-8)
amod(border-10, southern-9)
nmod(wall-6, border-10)
cc(build-3, and-12)
nsubj(pay-15, Mexico-13)
aux(pay-15, will-14)
conj(build-3, pay-15)
case(wall-18, for-16)
det(wall-18, the-17)
nmod(pay-15, wall-18)
```

原句：We will build wall and pay

TASK 4(CONT'D)

- 原句的Bigram model: (we will), (will build), (build a)….
- 簡化後的句子的Bigram model: (we will), (will build), (build wall)….
- Bigram Model的Entropy算法請參考(2-a p.10)

TASK 4輸出格式

- 前十行為簡化後的句子
- 第十一行分別為簡化後bigram model的大小、H(X)、H(Y | X)和Entropy(X, Y)
- 第十二行為分別為簡化前bigram model的大小、H(X)、H(Y | X)和E(X, Y)

We will build wall and pay
…
…
…
1000000,14.6,18.2, 32.8
50000000,20,30,50

規則

- 可使用任何套件
- 嚴禁抄襲，違者以零分計算
- 最晚請在15:20把檔案上傳至空間，之後算一律算補交
- 補交期限到10/25 23:59