自然語言處理與文字探勘

陳縕儂＆教研處

「版權聲明頁」

本投影片已經獲得作者授權台灣人工智慧學校得以使用於教學用途，如需取得重製權以及公開傳輸權需要透過台灣人工智慧學校取得著作人同意；如果需要修改本投影片著作，則需要取得改作權；另外，如果有需要以光碟或紙本等實體的方式傳播，則需要取得人工智慧學校散佈權。

– 台灣人工智慧學校

課程內容

講師投影片
資料與投影片
影片播放列表
程式碼：~/courses-tpe/NLP

1. 詞向量應用
   - Clustering Visualization
   - Document Vector

Code / Data 放在 hub 中的 courses 內

- 為維護課程資料 , courses 中的檔案皆為 read-only, 如需修改請 cp 至自身的環境中
- 打開 terminal, 輸入
  - [台北班]

    cp -r courses-tpe/NLP/part3 <存放至本機的名稱>
  - [新竹班]

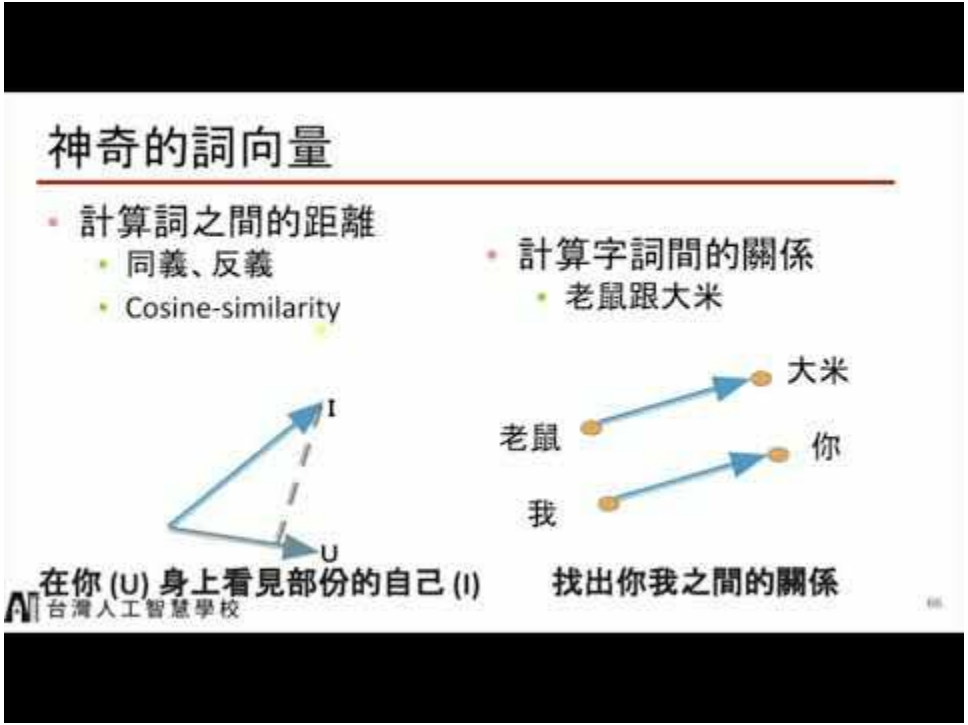
    cp -r courses-hsi/NLP/part3 <存放至本機的名稱>
  - [台中班]

    cp -r courses-txg/NLP/part3 <存放至本機的名稱>

程式實作

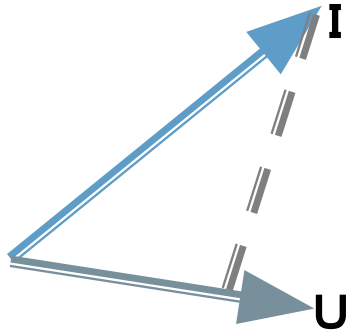
詞向量應用

詞向量應用


神奇的詞向量
計算詞之間的距離
同義、反義
Cosine-similarity
計算字詞間的關係
老鼠跟大米
I
U
大米
老鼠
你
我
在你 (U) 身上看見部份的自己 (I)
找出你我之間的關係
台灣人工智慧學校

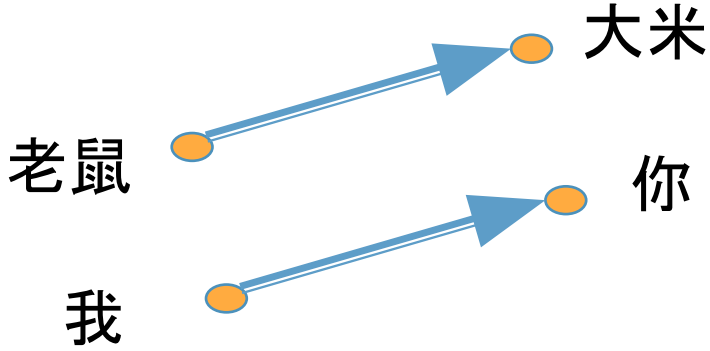
神奇的詞向量

- 計算詞之間的距離
  - 同義、反義
  - Cosine-similarity



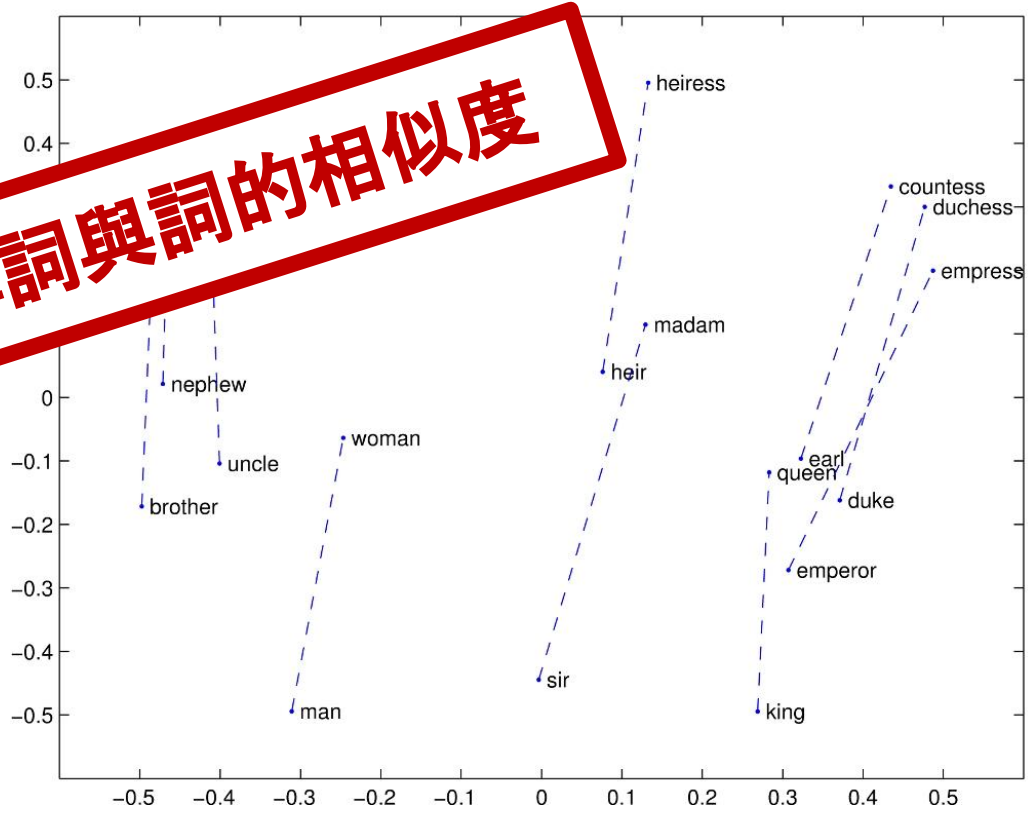
在你 (U) 身上看見部份的自己 (I)

- 計算字詞間的關係
  - 老鼠跟大米



找出你我之間的關係

From Stanford University Natural Language Processing Group GloVe Project



計算詞與詞的相似度

程式範例

similar

```
model.wv.most_similar('KMT')
```

```
[('國民黨', 0.6721848249435425),
 ('DPP', 0.645309567451477),
 ('kmt', 0.6443658471107483),
 ('dpp', 0.60954749584198),
 ('民進黨', 0.5811234712600708),
 ('執政', 0.5784410238265991),
 ('在野黨', 0.5460374355316162),
 ('執政黨', 0.5453152656555176),
 ('阿扁', 0.5433717370033264),
 ('新黨', 0.5219364762306213)]
```

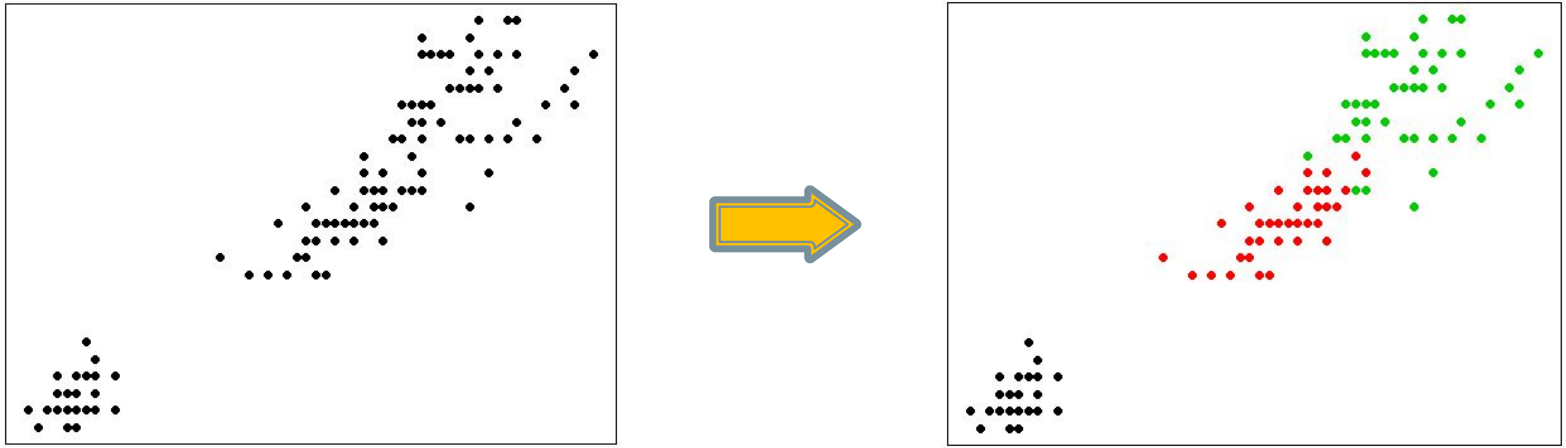
```
model.wv.most_similar(positive=['KMT', '綠吱'], negative=['DPP'])
```

```
[('外省人', 0.4196818470954895),
 ('賈國賊', 0.4115854799747467),
 ('周處', 0.4027564525604248),
 ('大話', 0.3970697224140167),
 ('英雄難過', 0.38398563861846924),
 ('外省', 0.3826029896736145),
 ('憤青', 0.3802189826965332),
 ('先祖', 0.37779176235198975),
 ('同表', 0.3775332570075989),
 ('禮智信', 0.37636640667915344)]
```

在字詞向量化後
相似的字詞之間距離較相近

將相似的詞做分群 (Clustering)

k-means clustering 資料分類



k-means 程式範例與結果

- 用 word vectors 的結果去做 kmeans clustering (sklearn)
  - 挑選 top n 字詞

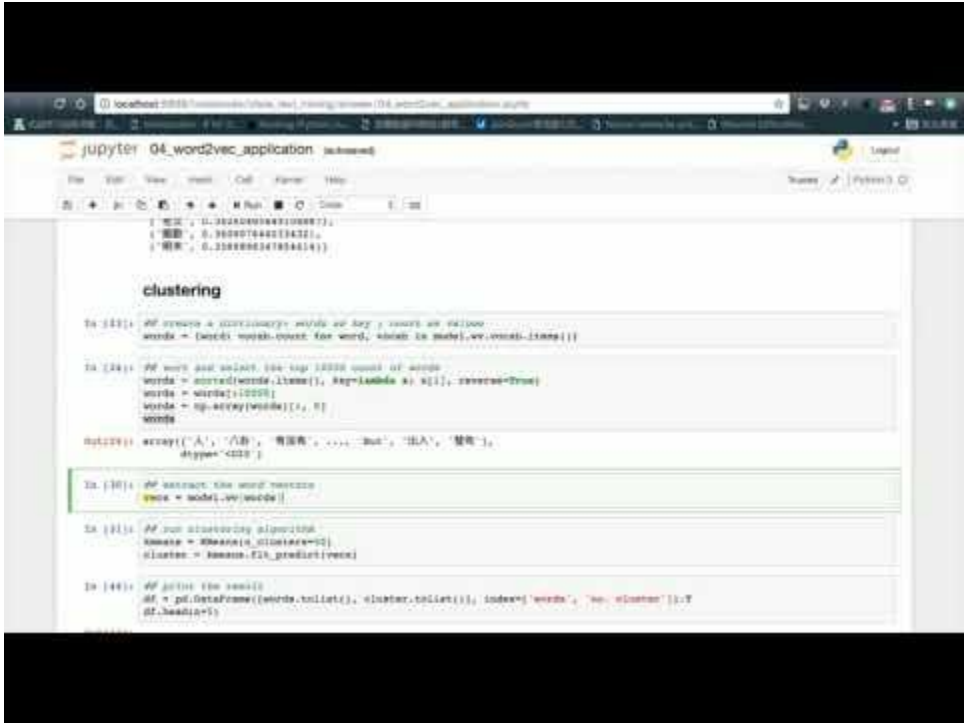
```
kmeans = KMeans(n_clusters=50)
cluster = kmeans.fit_predict(vecs)
```

	0	1	2	3	4	5	6	7	8	9	...	40	41	42	43	44	45	46	47	48	49
0	喝	比賽	錢	說	年	站	機車	館長	變成	未滿	...	問卦	新聞	最近	歲	小	穿	禁止	陳	字	iPhone
1	飲料	戰	賺	知道	世界	捷運	警察	神	出現	之內	...	推	影片	今天	長	老婆	衣服	需要	表示	被刪	手機
2	泡	大連	花	真的	新	搭	路□	最強	發生	問板	...	發文	討論	後	分	可愛	一件	目前	醫院	寫	遊戲
3	咖啡	世	有錢	覺得	當時	坐	騎	王	抓	繁體中文	...	版	網路	題	運動	姆	T	無法	規定	照片	電腦
4	茶	實力	賺錢	看到	成為	飛機	路	兄弟	造成	行文	...	看板	內容	前	度	讚	內褲	使用	單位	卡	科技
5	酒	一場	房子	好像	第一	鐵	一台	魔法	事件	の	...	兩篇	廣告	已經	年紀	名字	戴	已	人員	找到	技術
6	奶茶	輸	收	有人	幾年	位置	停	天堂	殺	い	...	板規	電視	小弟	比例	正	t	所有	報告	FB	設計
7	菸	棒球	每年	感覺	當年	公車	開車	強	嚴重	な	...	板	媒體	安安	身材	X	穿著	發展	黃	資料	系統
8	一杯	選手	銀行	發現	時代	線	危險	斬	導致	ん	...	文	直播	月	機率	年輕	頭髮	經濟	法	一張	蘋果
9	喝酒	騎士	投資	這種	最大	司機	車	英雄	自殺	る	...	文章	記者	本魯	距離	金城武	顏色	情況	未	圖	功能
10	牛奶	中華	支出	喜歡	曾經	高鐵	汽車	練	保護	e	...	PTT	網路□	小魯	秒	C	褲子	影響	同意	貼	軟體
11	杯	聯盟	金	算	加入	票	撞	宇宙	不斷	っ	...	ptt	查	一天	速度	腿	潮潮	方式	申請	電話	潮
12	戒	冠軍	收入	以前	全	火車	騎車	大師	消失	に	...	廢文	網站	幾天	肥	美	制服	要求	提出	臉書	自動

程式練習時間

- 04_word2vec_application.ipynb
  - 嘗試玩文字向量相似度 , e.g. [爸爸, 媽媽] 等於 [老公, ?]
  - 執行 kmeans 程式 , 並嘗試調整羣數

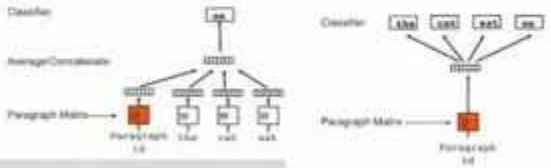
程式解說



Doc2Vec

Doc2Vec

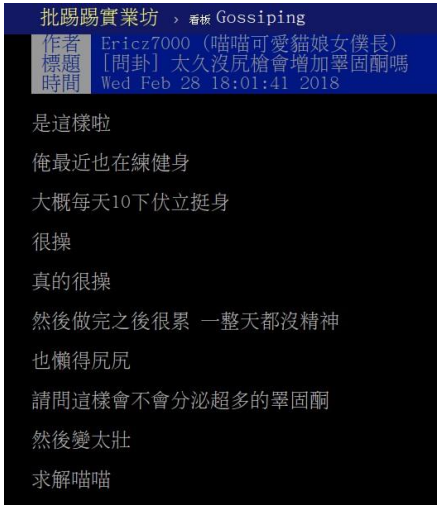
- 文章向量可以透過文字向量的平均處理作代表，但仍然忽略了單詞之間的排列順序對情感分析的影響
- Doc2Vec method，除增加段落向量，方法幾乎等同於word2vec
- Distributed Memory (DM) & Distributed Bag of Words (DBOW)



台灣人工智慧學校

詞向量轉換成文本向量

- 文章內每個詞向量加總取平均後，可以用來代表每篇文章的屬性 .



文章

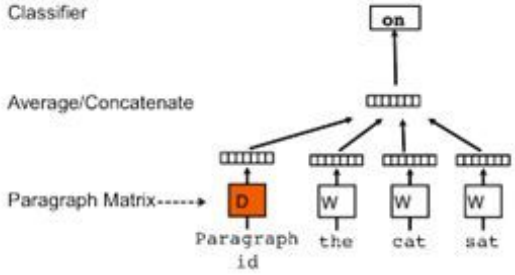
詞向量

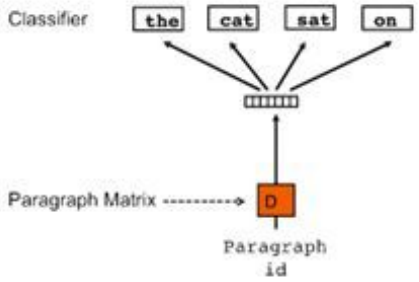
撕心裂肺	0.250113823	0.288580771	-0.349648024
常叫	0.581631128	-0.168796440	-0.617956259
范玄珍說	0.059279153	-0.474639349	-0.236295720
駐台	-0.376782237	-0.797129303	0.651099000
銀枝則	-0.605435689	-0.069144042	0.080284280
麗莉	0.471099780	-0.551091980	-0.321278178
縱	0.244521769	-0.053195253	0.122502663
節儉	0.435786642	-0.150547734	-0.642815039

A3930	0.050570641	-0.1578603	0.06377170
A3931	0.075977638	-0.1898975	0.06608932
A3932	0.052882687	-0.1672281	0.09590967
A3933	0.046331780	-0.1613533	0.09096074
A3934	0.073454491	-0.1611603	0.11529204

Doc2Vec

- 文章向量可以透過文字向量的平均處理作代表，但仍然忽略了單詞之間的排列順序對情感分析的影響
- Doc2Vec method，除增加段落向量，方法幾乎等同於 word2vec
- Distributed Memory (DM) & Distributed Bag of Words (DBOW)





Doc2Vec 程式範例

```
sentence_list = []

for i, l in enumerate(data):
    sentence_list.append(doc2vec.LabeledSentence(words=l, tags=[str(i)]))
```

```
/usr/local/lib/python3.5/dist-packages/ipykernel_launcher.py:4: Deprecation
ence` (Class will be removed in 4.0.0, use TaggedDocument instead).
  after removing the cwd from sys.path.
```

```
model = Doc2Vec(size=256, min_count=5, window=4, workers=10)
```

```
model.build_vocab(sentence_list)
```

程式練習時間

- 05_document_vector.ipynb
  - 練習提取文章的 word vector 並取平均
  - 使用 Doc2Vec 訓練文章向量

程式解說