文本分析與程式設計
Week03

本課程由卓騰語言科技贊助

取特徵詞

- 特徵詞可以做為代表文本內容的一種參考維度。換句話說，可以透過特徵詞來理解文本想傳達的主要概念。

- 上週的課程提到取特徵詞的方法有三種
  - TF-IDF
  - 詞性 (名詞 /動詞)
  - 「人、事、時、地、物」取特徵詞
- 這週會講解關於如何以「人、事、時、地、物」來取特徵詞

取特徵詞

- 為什麼要使用「人、事、時、地、物」來取特徵詞呢？
  - 如果有多篇文章中提及的「人」有大量的重覆，那麼我們大概可以推測它大概是在討論類似的主題。同樣的道理，同樣的「事、時、地、物」也可以做為分類文本時的依據。

取特徵詞

- 地點，在前次課程中已經有介紹了articut.getLocationStemLIST() 這個語法來幫助我們取出地點
- 週課程中，會介紹如何取出：

類別	範例語法
人	articut.getPersonLIST(baseballResultDICT)
事	articut.parse(baseballSTR, userDefinedDictFILE="./mixedDICT.json", level = "lv3")["event"]

取特徵詞

- 這週課程中，會介紹如何取出：

類別	範例語法
時	articut.parse(baseballSTR, userDefinedDictFILE="./mixedDICT.json", level = "lv3")["time"])
地	articut.getLocationStemLIST(baseballResultDICT)
物	articut.getNounStemLIST(baseballResultDICT)

課程目標

- 學習如何使用取出人事時地物的內容
- 知道為什麼要取出人事時地物的特徵詞

- 本課程使用的文本是從
- 1. https://news-taiwan.xyz/uncategorized/39053.html
- 2. https://www.ctwant.com/article/111388 (有經過編輯)

課程準備

- 以下的練習，我們會繼續使用棒球和籃球的兩篇新聞

從文本中抽詞

從文本中抽詞 -- 「人」

- 可以使用 articut.getPersonLIST() 來處理

```
baseballPeopleLIST =
articut.getPersonLIST(baseballResultDICT)
print(wordExtractor(baseballPeopleLIST))

basketballPeopleLIST =
articut.getPersonLIST(basketballResultDICT)
print(wordExtractor(basketballPeopleLIST))
```

棒球執行結果
['他', '巴斯', '庫爾帕', '康福托', '自己', '麥尼爾']

籃球執行結果
['保羅', '克拉克森', '康利', '戈貝爾', '米契爾']

課間練習1

- 以下有一篇娛樂新聞，請你利用 `articut.getPersonLIST()` 來找這篇新聞的人物。請問裡面有哪些人物呢？

- （中央社記者王心妤台北2日電）公視「勇者動畫系列」改編自台灣漫畫家黃色書刊作品，為公視動畫元年打頭陣，9月1日將上架Netflix，超過190個國家能看見，藝人黃子佼、吳慷仁、劉冠廷、孫可芳都獻聲。「勇者動畫系列」改編自以諷刺漫畫聞名的漫畫家黃色書刊2016年推出的網路連載漫畫「勇者系列」，目前已累積逾960話，不只有龐大世界觀，也用角色的刻板印象反諷社會。公視今天舉辦線上記者會，漫畫家黃色書刊、製作人王尉修、導演楊子霆、配樂師張衛帆，以及藝人黃子佼、旺福小民與Mami、美秀集團鍵盤手冠佑、鼓鼓呂思緯、小魏魏嘉瑩及黃奕儒都分享心得。對於是否擔心觀眾對漫畫變動畫的迴響，黃色書刊表示，因劇情仍照著漫畫進行，觀眾還是能夠有思考的空間。他也指出，漫畫被改編像是夢想實現，「這是很感動的，當動畫團隊來找我，我馬上就答應了。」他也特別創作全新角色「忠誠勇者」，將會起到貫穿全作的效果。黃子佼表示，當時僅跟團隊聊了幾分鐘就決定加入，為了配合角色「老魔王」的個性，特別壓低聲音並放慢語速。已經看過全作的他表示，「沒有人是絕對的好人，或絕對的壞人，不能用種族或者長相的分類去判斷，很呼應現實世界。」為「勇者動畫系列」創作歌曲「今世世界紀錄」的旺福樂團成員小民表示，腦袋在創作過程都是動畫情節，用熱血沸騰的心情完成創作，負責演唱的Mami則說，歌曲裡的龐克精神，能夠讓人找回初心。美秀集團鍵盤手冠佑的插曲「不能重生的冒險」，則抱持「雖然做了不一定成功，但是不做一定不會成功」的心態。公視「勇者動畫系列」將於4日起，每週日晚上10時在公視、公視+播出；8月1日則在myVideo；9月1日LINE TV、Netflix上架。（編輯：陳政偉）1100702

- 本文取自於 https://www.cna.com.tw/news/amov/202107020302.aspx

課間練習2

- 「事」，在這裡指的是「事件」。請和同學討論看看，一個事件應該要包含哪些內容呢？

從文本中抽詞 -- 「事」

- 人、時、地和物由articut 都可以直接取出相關的特徵詞
- 「事」，在這裡指的是「事件」。一個「事件」由「涉及的人/物」加上「動作」構成。

從文本中抽詞 -- 「事」

- 可以使用 articut.parse() lv3 的「語意分析」能力來取得事件
- 之前操作 lv1/lv2 一樣，透過 .parse() 的函式，同時傳入 mixedDICT.json 的字典檔，這次設定為 "lv3"。
- 並將最後計算後的結果取出 ["event"] 的值

```
baseballEventLIST = articut.parse(baseballSTR,
userDefinedDictFILE="./mixedDICT.json", level =
"lv3")["event"]
basketballEventLIST = articut.parse(basketballSTR,
userDefinedDictFILE="./mixedDICT.json", level =
"lv3")["event"]
```

從文本中抽詞 -- 「事」

- 棒球類文本輸出的結果如下

[[], ['此戰', '使用'], ['上陣', '打線'], ['壓制', '打線'], '\n', [], ['讓', '球隊'], ['帶著', '球隊'], ['進入', '優勢'], ['推出', '登板'], ['關門', '登板'], '\n', [], [], ['讓', '追平比數'], [], [], [], '\n', ['形成', '局面'], ['福', '輪到'], ['上場', '康福托'], ['打擊', '康福托'], [], ['投了', '角滑球'], ['眼', '看'], '\n', ['靠近', '自己'], ['伸進', '手'], ['碰觸', '他'], ['他', '到'], ['隨', '庫爾帕'], ['判定', '庫爾帕'], '\n', [], ['讓', '分'], ['拿下', '分'], ['再見', '分'], ['贏得', '勝利']]

- 從棒球比賽的文本裡可以看到的是有「帶著 - 球隊」、「進入 - 優勢」...等等的事件發生。甚至可以看到最後是「贏得 - 勝利」而可以推測該句的主角在最後應該是贏了比賽。

從文本中抽詞 --「事」

- 籃球類文本輸出的結果如下

[[], ['錯失', '勝利'], ['開始', '攻勢'], ['打出', '攻勢'], ['比', '攻勢'], '\n', ['力圖', '康利'], ['追分', '康利'], [], ['讀秒', '階段'], ['上籃', '階段'], ['得手', '階段'], ['罰', '中'], ['留給', '機會'], ['追平', '機會'], '\n', ['造成', '米契爾'], ['犯規', '米契爾'], ['罰', '中'], ['隨', '勝利'], ['用', '勝利'], ['罰', '勝利'], ['鎖定', '勝利'], ['狂轟', '米契爾'], ['助攻', '米契爾'], ['得分', '次'], [], '\n', [], [], [], ['助攻', '康利'], [], ['人', '合計'], [], '\n', [], []]

- 從籃球比賽的文本裡，則可以看到「錯失-勝利」、「狂轟-米契爾」、「助攻-米契爾」...等等事件。

課程練習3

- 請使用課程練習1提及的那則娛樂新聞以及裡面的來分析裡面有哪些事件。

從文本中抽詞 -- 「時」

- 「時間」資訊也是屬於 lv3 語意分析的範疇
- 因此操作上和前述的「事件」一樣，只在最後取出的是 ["time"] 而不是 ["event"] 而已。

```
baseballTimeLIST = articut.parse(baseballSTR,
userDefinedDictFILE="./mixedDICT.json", level =
"lv3")["time"]
print(baseballTimeLIST)

basketballTimeLIST = articut.parse(basketballSTR,
userDefinedDictFILE="./mixedDICT.json", level =
"lv3")["time"]
print(basketballTimeLIST)
```

從文本中抽詞 -- 「時」

- 籃球類文本輸出的結果如下

[[{'absolute': False, 'datetime': '2021-07-04 22:00:00', 'text': '昨晚', 'time_span': {'year': [2021, 2021], 'month': [7, 7], 'weekday': [7, 7], 'day': [4, 4], 'hour': [22, 2], 'minute': [0, 59], 'second': [0, 59], 'time_period': 'night'}}], [], [], '\n', [], [], [], [], [], '\n', [], [], [], [], [], [], '\n', [], [], [], [], [], [], '\n', [], []]

- 另一篇籃球比賽的文本裡，則「昨晚的紐約西區霸王之戰中...」中，有「昨晚」這一個表示時間的詞彙。Articut lv3 的 [“time”] 則算出以下的結果。一樣取出了「昨晚」並加以計算出它的時間在 “2021-05-06 22:00:00”

課間練習4

- 請使用課程練習1提及的那則娛樂新聞以及裡面的來分析裡面提及那些時間點。

從文本中抽詞 - 「地」一般地點

- Articut 自帶的 **.getLocationStemLIST()** 函式可以像前述取得人名列表一樣地操作，將文本中指涉「某種地方」或「位置」的詞彙抽出。

```
baseballlocLIST = articut.getLocationStemLIST(baseballResultDICT)
print(baseballlocLIST)

basketballlocLIST = articut.getLocationStemLIST(basketballResultDICT)
print(basketballlocLIST)
```

執行結果
['球帶內', '紐約']

籃球執行結果
['紐約', '西區']

從文本中抽詞 - 「地」景點

- Articut 也可以取得景點，就使用 articut.parse()
- 請使用下一頁的文章

- 本文取自於 https://yencheng0817.pixnet.net/blog/post/326445630

從文本中抽詞 - 「地」

```
inputSTR = '''
```
澎湖由本島以及周邊離島組成，以旅遊路線劃分本島，可以分成四大區域：馬公市區、北環、南環、以及澎湖機場以東的東環。至於澎湖離島分為四大海域：東海（鳥嶼、員貝、澎澎灘）、北海（吉貝、目斗嶼、險礁）、南海（七美、望安、虎井、桶盤）及南方四島國家公園（東嶼坪、西嶼坪、東吉嶼及西吉嶼）。熱門的七美、望安，是屬於南海四島，不要跟南方四島搞混囉！◈ 馬公市區：遊客住宿大多會選在馬公市區，尤其是中央老街周邊，美食及住宿選擇很多，離南海遊客中心也近。◈ 北環：有許多澎湖必去景點，包括跨海大橋、鯨魚洞、柱狀玄武岩、燈塔...等，通常會安排一整天的一日遊。◈ 南環：以沙灘及海邊景觀為主，山水沙灘傍晚極美。◈ 東環：最著名的景點是奎壁山摩西分海，還可以在隘門沙灘上喝咖啡享受悠閒夏日。★ 我會建議澎湖旅遊行程天數至少安排 4 天，其中 3 天在本島，跑北環線加上東環、南環景點，以及馬公市區的吃喝逛街。（下面會有更詳細的景點介紹）★ 行程中可以穿插 1 - 2 天跳島，視個人喜好安排離島海域。★ 先確認要去的離島是在哪一個海域，再分別到南海、北海、東海遊客中心搭船（三個海域的遊客中心地址都不同）。船班時間一天只能安排一個海域，下午可以回馬公市區吃喝。➤ 旅遊季節澎湖 4 - 9 月為旅遊旺季，冬天東北季風強勁，較不推薦冬天時前往旅行。個人推薦 4、5 月最適合，開始進入夏季，太陽也不會過於炙熱，還可以避開暑假的恐怖人潮。馬公市區◈ 市區景點以天后宮及周邊的中央老街為主，有許多美食及住宿選擇。一級古蹟澎湖天后宮是台灣歷史最悠久的媽祖廟。中央老街是澎湖最早發展的街道，古色古香的街道建築很適合漫步。四眼井旁的乾益堂中藥行已開業超過百年，來到這可以品嘗看看他們的藥膳蛋和豆干。北環北環線是澎湖經典旅遊路線，從馬公市區一路到最遠的西嶼燈塔（漁翁島燈塔），沿途有許多知名必去景點。下面依照從市區出發的經過順序介紹：◈ 後寮天堂路白沙鄉後寮村的天堂路，這幾年是越來越熱門的澎湖秘境景點。
```
'''
```

從文本中抽詞 - 「地」

- 請看範例程式

```
penghuLIST = articut.parse(inputSTR, openDataPlaceAccessBOOL = True)
```

執行結果

```
10
10 20 30
```

- 本文取自於 https://yencheng0817.pixnet.net/blog/post/326445630

課間練習5

- 請使用課程練習1提及的那則娛樂新聞以及裡面的內容來分析裡面有哪些地點，也利用取得景點的語法，來看看裡面是否有特殊的景點。

從文本中抽詞 - 「物」

- 從文本中抽取其意義指「物品」的詞彙組
- 這個功能其實和前一週提到的「抽出名詞」是一樣的。

```
baseballNounLIST = articut.getNounStemLIST(baseballResultDICT)
print(baseballNounLIST)

basketballNounLIST = articut.getNounStemLIST(basketballResultDICT)
print(basketballNounLIST)
```

棒球執行結果
['主審', '優勢', '分', '勝利', '安打', '局面', '意識地', '手', '打線', '打者', '投手', '此戰', '比賽', '滿壘', '球', '球隊', '登板', '眼', '角滑球', '觸身球', '身肘', '身體', '車輪戰', '追平比數', '這球', '陽春砲', '領先']
籃球執行結果
['中', '人', '勝利', '單場', '場', '布克', '延長賽', '戰', '攻勢', '機會', '次', '波格丹諾維奇', '籃板', '階段', '霸王']

課間練習6

- 請使用課程練習1提及的那則娛樂新聞以及裡面的來分析裡面有哪些「物」。

討論

- 剛剛在不同的練習中練習如何取得人事時地物，不過目前練習到現在，大家覺得電腦幫你挑出人事時地物，請問你覺得挑出這些特徵詞對分類文本有幫助嗎？會怎麼幫助你呢？

人事時地物特徵詞應用brainstorm

- 參考內容

- 看看洗錢文章中可以直接取得人名，接下來可以分析誰是犯人，誰是檢察官，接下來取得的人名可以繼續做後續分析
- 如果有好大一篇故事，例如水滸傳，可以取得每個章節的事件，藉此可以用最短時間整理每個章節的大意

作業:遠距教學新聞的特徵詞抽取

作業敘述:

- 在處理股市文本任務中，我們傾向將文本分成2種:

1. 描述跌的文本
2. 描述漲的文本

- 而這次的任務裡，我們從聯合新聞網找了近期的股市新聞(20篇)，希望可以透過我們學到的抽取特徵詞的技術，幫我們順利辨認出有關於敘述股市漲的文本是那些，讓我們直接開始吧!

作業:遠距教學新聞的特徵詞抽取

- 任務一: 思考時間
- 從文本中我們可以觀察出以下幾個關鍵，首先我們會關心這是在說哪一支股票，第二部分我們會好奇這支股票是漲還是跌。
- 那我們是怎麼判斷文章是再說漲還是跌呢? 還明顯這時候我們無法依賴單純斷詞後的詞頻。在我們解決地個問題之前，我們先用我們的語感來觀察資料，你覺得那些句子透露出一篇文章的漲跌線索呢?
- 文本在下一頁

任務1文本

- 元太電子紙商機起飛 股價挑戰歷史新高
- 電子紙大廠元太（8069）因市場看好電子紙閱讀器與筆記本，以及電子紙標籤（ESL）商機，近期股價持續走揚，今（6）日盤中高點來到84.5元，上漲7元、漲幅達9%，近期有望挑戰歷史最高價85.6元。 元太今年前5月累計合併營收達71億元，年增26.5%，是自2017年轉型100%電子紙製造商後的同期新高。...
- 原文連結

作業:遠距教學新聞的特徵詞抽取

- 任務二之一: 抽取需要的特徵詞
- 在這個任務中，請你根據所學過的抽取特徵詞的方法，包含運用TFIDF值、抽取名詞、抽取動詞、抽取事件這四種技術，分別以第一則新聞去嘗試，之後綜合比較，你覺得要用哪個方法來判斷一篇股市新聞是漲是跌比較好呢?

作業:遠距教學新聞的特徵詞抽取

- 任務二之二: 根據詞性自定義程式
- 在這個任務中，我們希望自己透過對詞性標記的觀察，留下標記為UserDefined, ACTION, ENTITY, VerbP，並且只留下句子(個人定義句子為有Entity/UserDefined跟Action)，希望透過這樣子的方式我們可以過濾目前不需要的元素，但是又保持句子的概念，如此一來，當我們發現“股價”跟“走揚”出現在同一個句子時，我們就可以比較放心認定他是一個有關股市上漲的句子了。

作業:遠距教學新聞的特徵詞抽取

- 任務二之二: 根據詞性自定義程式產出範例

```
[['市場', '看好', '電子紙', '閱讀器', '筆記本'],
 ['股價', '持續', '走揚'],
 ['日盤', '高點', '來到'],
 ['漲幅', '達', '9%'],
 ['挑戰', '歷史'],
 ['元太', '累計', '合併', '營收', '達'],
 ['增', '26.5%'],
 ['轉型', '100%', '電子紙', '製造', '商', '同期'],
 ['元太', '指出'],
 ['疫情', '加速', '電子貨', '架標籤', '裝', '機潮'],
 ['公眾', '顯示器', '物流', '應用', '同步', '升溫'],
 ['讓', '元太', '客戶端', '需求', '成長'],
 ['時序', '進入'],
 ['市場', '迎來', '解封', '開學'],
 ['元太', '擴增', '全新', '彩色', '電子', '紙技術', '產'],
```

作業:遠距教學新聞的特徵詞抽取

- 任務二之三:探索articut.getContentWordLIST()這個功能
- 這是一個超便利小工具，如果前面對你而言有點太難，在articut裡面也有一個類似的小工具，我們來嘗試看看吧!!
- articut.getContentWordLIST() 回傳結果範例:

```
[[(14, 17, '電子紙'), (46, 48, '大廠'), (76, 78, '元太')],
 [],
 [],
 [],
 [(43, 45, '市場'),
  (76, 78, '看好'),
  (106, 109, '電子紙'),
  (138, 141, '閱讀器'),
  (208, 211, '筆記本')],
```

作業:遠距教學新聞的特徵詞抽取

- 任務二之四: 思考時間

請回想一下之前的思考活動，我們認定有一些句子是能有句於我們理解一篇文章是描述漲還是跌，我們可以試著將觀察是那些字詞讓我們有這樣的感受呢? 例如在第一句我們可以發現是**股價**和**走揚**，讓我們有這樣的認知，那我們是不是能夠建立一個LIST在其中，我們把有關股價上漲的字詞都放進去，如果一個有出現這些那我們就可以暫時認定文章跟這方面有關係，如果這樣的句子越多，我們就越能肯定這個文章跟股市上漲有關。

作業:遠距教學新聞的特徵詞抽取

- 任務三: 開始判斷文本
- 在這個任務中，我們開始要將我們在思考時間所累積的想法實踐出來，我們首先要先藉由人工觀察，建出一個有關於描述股市漲的LIST，之後將它和新聞去進行比對以得到結果

作業:遠距教學新聞的特徵詞抽取

- 範例結果

Title	Content	go_up_score
元太電子紙商機起飛 股價挑戰歷史新高	電子紙大廠元太（8069）因市場看好電子紙閱讀器與筆記本，以及電子紙標籤（ESL）商機，近期...	0.256
台股再登新高後壓回 航海王權證最熱門	台股今（6）日早盤在傳產與金融族群支撐盤勢，多頭指標—海運股買盤持續湧入下，指數一度登上「...	0.087
聯詠6月、Q2營收同創新高 上半年應可賺回兩個股本	驅動晶片廠聯詠（3034）今（6）日公告6月合併營收續升至115.8億元，較上月微增1.2%...	0.469
台股衝關萬八終場收跌6.26點 三大法人賣超33.09億	台股今（6）日開高後一度衝上18,008.37點，再創歷史新高，首度越過18,000點大關...	0.207
高端起落坑殺人 立委籲加嚴炒股查緝	「股票炒半天，大起大落、暴漲暴跌，被有心人賺走錢，散戶什麼都不剩！」台灣民眾黨立委張其祿說...	0.023
除權息遞延衝擊！0050將除息0.3元、殖利率0.2%	「國民ETF」元大台灣50（0050）今年下半年的配息出爐，元大投信公告，0050每單位擬發...	0.114
陽明5月獲利240.35億元 EPS再創新高達3.15元	陽明海運（2609）今（6日）公布5月獲利自結獲利，再度寫下新高記錄，一個月輕鬆賺進百億元。...	0.188
國巨第二季營收季增逾16% 並較去年成長逾一倍	被動元件龍頭國巨（2327）今（6）日公布6月自結合併營收為95.05億元，單月營收較上月增...	0.333

作業:遠距教學新聞的特徵詞抽取

小結:

- 從以上的任務，雖然只是一個小嘗試，但我們可以發現，分數較低的文本的確就是跟股票漲價比較不相關，如下所示，一則是有關法條，一則是有關個人投資的失敗，一則是有關於外商投資策略，

Title	Content	go_up_score
高端起落坑殺人 立委籲加嚴炒股查緝	「股票炒半天，大起大落、暴漲暴跌，被有心人賺走錢，散戶什麼都不剩！」台灣民眾黨立委張其祿說...	0.023
他當沖狂賺幾萬塊「想玩大一點」 交易金額提高300萬卻GG了	台股今（6）日早盤指數一度登上「萬八」大關，再創歷史新高，而近年台股創高也引起許多投資客走上...	0.037
台積長線股價 外資喊千元	外電報導，英特爾將包下台積電（2330）3奈米產能，外資摩根大通、華興資本及瑞士信貸昨（6）...	0.044

作業:遠距教學新聞的特徵詞抽取

- 結語:
- 當然這個方法還是有不少缺陷，或許我們可以設計另一個功能判斷他是跌的分數，然後去和漲的分數進行比較，又或是有更好的計算分數的方式。那這部分就交給有興趣延伸的人更進一步探討囉! 加油!