文本前處理：Jieba 斷詞、詞性標記與句法學

台師大通識教育課程

文本分析與程式設計

授課：卓騰語言科技 _ PeterWolf

前處理

文本處理很像一種「漁業」

- 網眼太大：只撈得到大魚和大型垃圾
- 網眼太小：大小魚和垃圾都一起撈起來
- 適當處理手法：
  - 斷句 (a.k.a. 去除標點符號)
  - 斷詞 (今天以 "Jieba" 斷詞為例)
  - 去除語意意含較低的詞彙 (HOW?)
  - TF-IDF 就是 "keyword (關鍵詞)" 嗎？

斷句:

Application security is hard...when it’s a separate process. With GitLab, application security testing is built into the CI/CD process. Every merge request is scanned for vulnerabilities in your code and that of its dependencies.

「不過SBL新球季改成單洋將，又限制身高，這樣讓我們這種4、5號位置的球員發揮空間較大，我當然希望可以多打一點..」范士恩說，「其實到現在我仍在學習，像是我的大學長周柏臣，感覺他在打內線的時候都很輕鬆。」

一名男子今天清晨下大夜班，騎機車從中西區樹林街西往東，清晨6時 10 分在南門路口遇上交通管制，繞了3條路都繞不出南門路管制線，被困在府前路、南門路與健康路包圍區域內，找不到回家的路，他詢問路口管制交通的許姓警員。

陳先生嗎我是你兒子綁匪在我手上你準備三百萬來吧兒子啊乖叫聲爸爸來聽聽哎呀我說你呀你沒事綁綁匪幹嘛呢人家也是出來討生活的快把他給放了吧喂喂喂你掛我電話

斷詞要做什麼：CWS (Chinese Word Segmentation)

#觀察資料

```
["潤寅負責人楊文虎",
  "王音之夫婦涉嫌自民國 99 年 8 月起",
  "向 9 家銀行詐貸新台幣 386 億 2,718 萬餘元",
  "台北地檢署今年1月間依違反銀行法等罪嫌起訴楊文虎",
  "王音之等 36 人",
  "檢方另查出",
  "楊文虎等人涉向3銀行詐貸 86 億餘元",
  "5 月間追加起訴楊文虎等 15 人及潤寅等 5 家公司"
]
```

#期待結果：

```
["潤寅/負責人/楊文虎",
  "王音之/夫婦/涉嫌/自/民國/ /99/ /年/ /8/ /月/起",
  "向/ /9/ /家/銀行/詐貸/新台幣/ /386/ /億/ /2,718/ /萬/餘/元",
  "台北/地檢署/今年/1/月/間/依/違反/銀行法/等/罪嫌/起訴/楊文虎",
  "王音之/等/ 36/ /人",
  "檢方/另/查出",
  "楊文虎/等/人/涉/向/3/銀行/詐貸/ /86/ /億/餘/元",
  "5/ /月/間/追加/起訴/楊文虎/等/ /15/ /人/及/潤寅/等/ /5/ /家/公司"
]
```

怎麼做？ **n-gram** 嗎？

N-Gram

這個星期日本想往後山藥師佛寺去世人罕至處想一想自己的人生

uni-gram:

這/個/星/期/日/本/想/往/後/山/藥/師/佛/寺/去/世/人/罕/至/處/想/一/想/自/己/的/人/生

bi-gram:

這個/個星/星期/期日/日本/本想/想往/往後/後山/山藥/藥師/師佛/佛寺/寺去/去世/世人/人罕/罕至/至處/處想/想一/一想/想自/自己/己的/的人/人生

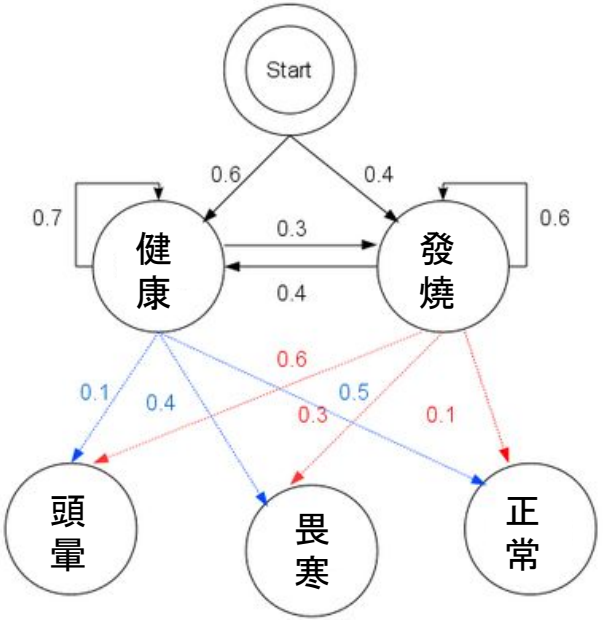
tri-gram:

...

Jieba 結巴斷詞

https://github.com/fxsjy/jieba

- 背景:
  - 1998 人民日報
  - 新詞基於 HMM (隱馬可夫模型 Hidden Markov Model)
    - [抗击新冠肺炎疫情斗争取得重大战略成果]
    - [抗擊新冠肺炎疫情鬥爭取得重大戰略成果]
- 重點:
  - TF-IDF
- 使用方法:
  - 安裝:pip3 install jieba
  - 使用:見文件



安装 jieba https://github.com/fxsjy/jieba#安装说明

安装说明

代码对 Python 2/3 均兼容

- 全自动安装：`easy_install jieba` 或者 `pip install jieba` / `pip3 install jieba`
- 半自动安装：先下载 http://pypi.python.org/pypi/jieba/ ，解压后运行 `python setup.py install`
- 手动安装：将 jieba 目录放置于当前目录或者 site-packages 目录
- 通过 `import jieba` 来引用
- 如果需要使用paddle模式下的分词和词性标注功能，请先安装paddlepaddle-tiny，`pip install paddlepaddle-tiny==1.6.1`。

使用 jieba https://github.com/fxsjy/jieba#主要功能

代码示例

```
# encoding=utf-8
import jieba

jieba.enable_paddle()# 启动paddle模式。 0.40版之后开始支持，早期版本不支持
strs=["我来到北京清华大学","乒乓球拍卖完了","中国科学技术大学"]
for str in strs:
    seg_list = jieba.cut(str,use_paddle=True) # 使用paddle模式
    print("Paddle Mode: " + '/'.join(list(seg_list)))

seg_list = jieba.cut("我来到北京清华大学", cut_all=True)
print("Full Mode: " + "/ ".join(seg_list))  # 全模式

seg_list = jieba.cut("我来到北京清华大学", cut_all=False)
print("Default Mode: " + "/ ".join(seg_list))  # 精确模式

seg_list = jieba.cut("他来到了网易杭研大厦")  # 默认是精确模式
print(", ".join(seg_list))

seg_list = jieba.cut_for_search("小明硕士毕业于中国科学院计算所，后在日本京都大学深造")  # 搜索引擎
print(", ".join(seg_list))
```

實際操作

```
sentenceLIST = [
"潤寅負責人楊文虎",
"王音之夫婦涉嫌自民國 99 年 8 月起",
"向 9 家銀行詐貸新台幣 386 億 2,718 萬餘元",
"台北地檢署今年1月間依違反銀行法等罪嫌起訴楊文虎",
"王音之等 36 人",
"檢方另查出",
"楊文虎等人涉向3銀行詐貸 86 億餘元",
"5 月間追加起訴楊文虎等 15 人及潤寅等 5 家公司"]
```

```
import jieba
resultLIST = []
for s in sentenceLIST:
    resultLIST.append("/".join(jieba.cut(s)))

print(resultLIST)
```

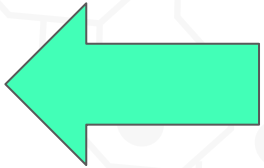
```
['潤寅/負責人/楊/文虎',
'王音/之夫婦/涉嫌/自民國/ /99/ /年/ /8/ /月/起',
'向/ /9/ /家銀行/詐貸/新/台幣/ /386/ /億/ /2/,/718/ /萬餘元',
'台北/地檢署/今年/1/月間/依違/反銀行法/等/罪嫌/起訴/楊/文虎', '王音/之/等/ /36/ /人',
'檢方/另/查出',
'楊/文虎/等/人涉/向/3/銀行/詐貸/ /86/ /億餘元',
'5/ /月間/追加/起訴/楊/文虎/等/ /15/ /人/及潤寅/等/ /5/ /家/公司']
```

TF-IDF: 關鍵字？

許多網站或是文件都會把 "TF-IDF" 和「關鍵字/詞」擺在一起:
https://www.google.com/search?channel=fs&q=TF-IDF+關鍵

TF: Term Frequency (這個字在本文件裡出現的頻率)
IDF: Inversed Document Frequency (這個字在所有文件裡出現的頻率)

```
if TF-IDF == 關鍵字:
    print("對啊！")
else:
    print("好像不太一樣！")
```



def 關鍵字_人的直覺():

具有代表性意義的詞彙

def TF-IDF():

和別的文件相比最不一樣的字串

停用詞 (stop word)

https://en.wikipedia.org/wiki/Stop_word

#定義 :

停用詞大致分為兩類。一類是人類語言中包含的功能詞..與其他詞相比，功能詞沒有什麼實際含義，比如'the'、'is'、'at'、'which'、'on'等...。另一類詞包括詞彙詞，比如'want'等，這些詞應用十分廣泛，但是對這樣的詞搜尋引擎無法保證能夠給出真正相關的搜索結果，難以幫助縮小搜索範圍，同時還會降低搜索的效率，所以通常會把這些詞從問題中移去，從而提高搜索性能。

#思考 :

對中文而言，停用詞應該具有什麼特徵？

1. 功能詞（詞性）
2. 短（單一字符「的」、「是」..）
3. 列個完整的停用詞表，對中文而言是最佳方案嗎？ https://www.google.com/search?channel=fs&q=停用詞表

中文裡，哪些類型的字是語意含量較低的？是「短」的字？還是「某些詞性類別」的字？

周天成/的/逆轉秀/不僅/讓/外媒/大呼/不可思議/，/與/香港/伍家朗/在/冠軍戰/的/最後/一球/也/被/BWF/選為/單月/最佳/好球/。/而/小天/連續/在/印尼/及/泰國/奪冠/，/積分/擠下/中國/好手/石宇奇/，/排名/升/至/世界/男單/第二/，/創/生涯/最佳/，/許多/外媒/也/看好/他/挾/這/股/氣勢/在/接下來/的/世錦賽/中/打出/佳績/。/

小結語

評價一下 Jieba 斷詞工具:

- 優點:
  - 快
  - 免費
- 缺點:
  - 斷詞錯誤率頗高
  - 語言受模型限制，即便做了簡繁轉換，但「語言」不是「編碼」而已
  - TF-IDF 的計算，排除「單字符的字詞」沒什麼道理
  - 詞性標記能力不佳，無轉品能力

Quiz:

課堂中已經演示了文本處理中，利用 Jieba 進行「斷詞工作」的幾個任務。試思考以下問題：

1. 斷詞系統的運作原理有哪幾種？Jieba 是屬於哪一種？
2. 「訓練文本」和「應用文本」的差異，是否會造成 Jieba 的表現不佳？
3. TF-IDF 是什麼意思？
4. TF-IDF 取出的「文件特徵字」是否就是「文件關鍵字」？
5. Stop word (停用字) 是什麼意思？
6. 「停用字」該採用「列舉詞條」方式處理，還是「依詞性判斷」？

Assignment:

1. 從課程 github repo 中把課程中提供的 week07 的目錄 git pull 下來。
2. 把 week07.py 改名為 **week07_你的學號_分組隊名.py**
3. 在 **week07_你的學號_分組隊名.py** 中，設計你的程式完成以下指定規格：
    a. 設計一 func() 名為 "text2cws(jsonFilePath)"，接受參數為一 .json 格式的檔案，並讀取 json 檔案中的 "BODY" 欄位的字串，加以「斷句」以後，使用 Jieba 斷詞將每個句子進行斷詞處理，回傳 值為一「斷詞處理後的列表」。
    b. 設計一 func() 名為 "termFreq(inputLIST)"，接受參數為列表，並依列表 內容的「字串元素」建立一字典 dict 型別的變數，將每個字串元素視為 key，整份文件中的，該字串元素出現的次數視為 value。
    c. 設計一程式進入點，透過前述 "text2cws()" 讀取 example/health/ 中所有檔案的 "BODY" 欄位的值，再透過 termFreq() 計算每個斷詞處理後的字串出現的次數。
    d. 同樣的步驟，再對 example/finance/ 中所有的檔案再處理一次。