迴圈控制及句法學

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

尋找句子的終點

對「語言」而言，不存在最長的句子！

任何句子都可再加上多個形容詞、副詞、子句來延長。

對「文本」而言，句子的長度則有限。

- 通常標點符號標誌著句子的終點 (e.g., 。，？！)
- 偶爾標點符號好像不存在，或不是終點 (e.g., \n, 12.03)
- 有些時候，文本裡「缺乏」標點符號 (e.g., 語音辨識)

實例說明:

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。」

通 常

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。

偶 爾

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

有時候

英文怎麼做:

#觀察資料

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.

#流程思考:

#要在所有的 ". " (一個點, 加一個空格) 之後切開句子。

```
inputLIST = inputSTR.split(". ")
```

#要在所有的 ". " 以及 "," 之後切開句子。

這表示我們有「一個以上」的條件, 要執行「一模一樣」的動作。

for 迴圈, 參上!

英文怎麼做：EN01.py

#要在所有的 ". " 以及 ", " 之後切開句子。

這表示我們有「一個以上」的條件，要執行「一模一樣」的動作。

```
for item in (". ", ", "):
    Do_Something_here
```

```
markLIST = [". ", ", "]
for index in range(0, 2):
    Do_Something_here
```

```
for item in (". ", ", "):
    inputSTR = inputSTR.replace(item, item+"<My_Cutting_Mark>")

inputLIST = inputSTR.split("<My_Cutting_Mark>")
print(inputLIST)
```

中文怎麼做：ZH01.py

#觀察資料

「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。」

#實作：

```
for item in ("「", "，", "；", "、", "...", "」", "。"):
    inputSTR = inputSTR.replace(item, item+"<My_Cutting_Mark>")


inputLIST = inputSTR.split("<My_Cutting_Mark>")
print(inputLIST)
```

中文怎麼做：ZH02.py

#觀察資料

「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。」

#實作：

```
for item in ("「", "，", "；", "、", "...", "」", "。", "："):
    inputSTR = inputSTR.replace(item, item+"<My_Cutting_Mark>")


inputLIST = inputSTR.split("<My_Cutting_Mark>")
print(inputLIST)
```

while 迴圈，參上！

中文怎麼做:

#觀察資料

所以希望可以借用學長姐的時光機😂拜託你們告訴我們一些經驗談吧！🙏🙏🙏（另外也想知道各位學長姐的未來目標！）

#實作:

http://www.unicode.org/reports/tr51/
http://www.unicode.org/Public/emoji/1.0//emoji-data.txt

https://gist.github.com/msenol86/44082269be46aa446ccda9d02202e523

中文怎麼做：ZH03.py

#觀察資料

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。

#實作：

```
inputSTR = inputSTR.replace("\n", "") #把斷行符「黏」回成一句
for item in ("「", "，", "；", "、", "…", "」", "。"):
    inputSTR = inputSTR.replace(item, item+"<My_Cutting_Mark>")

inputLIST = inputSTR.split("<My_Cutting_Mark>")
print(inputLIST)
```

中文怎麼做:

#觀察資料

我今天，買了一杯可不可
放在車箱裡面
想說等到目的地再喝

#實作:

```
inputSTR = inputSTR.replace("\n", "<My_Cutting_Mark>") #把斷行符號當成「斷句符號」
for item in ("「", "，", "、", "…", "」", "。"):
    inputSTR = inputSTR.replace(item, item+"<My_Cutting_Mark>")

inputLIST = inputSTR.split("<My_Cutting_Mark>")
print(inputLIST)
```

中文怎麼做:

#觀察資料

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

#思考:

對「語言」而言，不存在「最長」的句子，因此從「語言」的角度來說，硬要說這是「一句」也不是不行。

對「文本處理」而言，因為文本的基礎是「語言」，因此想要從「文本」去逆推「語言裡哪裡該斷句」是相當困難，且成效有限的。

這個問題可以透過大量的資料來建立斷句模型，但必需要知道「斷錯」是可預期的。

Quiz:

課堂中已經演示了文本處理中，遇到各種斷句符號以及沒有斷句符號時處理方法。試思考下列問題：

1. 「標點符號」的意義是否完全等同於「斷句符號」？
2. 「換行符號」的意義是否完全等同於「斷句符號」？
3. 文言文 (古文) 的文本是沒標點符號，也沒有換行符號的，
   那它有沒有斷句符號？
4. 所以...斷句處理和「句法」(syntax) 是否有關係？

尋找句子的終點

內建斷句標記

中文文字系統~~黏在一起~~

宋慶元初趙子直當國加朱文公為侍講文公欣然而至積誠感悟且編次講義以進寧宗喜令點句以來他日請問上曰宮中常讀之大要主求放心耳公因益推明其說曰坐下既知學問之要願勉強而力行之退謂其徒曰上可與為善若與得賢者輔道天下有望矣然

宋慶元初

趙子直當國

召朱文公為侍講

文公欣然而至

積誠感悟且編次講義以進

寧宗喜令點句以來

他日請問上曰宮中常讀之大要主求放心耳

公因益推明其說曰坐下既知學問之要

願勉強而力行

之謂其徒曰上可與為善

若與得賢者輔道天下有望矣然

Assignment:

1. 從課程 github repo 中把課程中提供的 week06 的目錄 git pull 下來。
2. 把 week06.py 改名為 **week06_你的學號.py**
3. 在 **week06_你的學號.py** 中，設計你的程式完成以下指定規格：
    a. 設計一 func() 名為 "jsonTextReader(jsonFilePath)"，接受參數為一 .json 格式的檔案，而 return 回傳值為 json 檔案中的 "text" 欄位的字串。
    b. 設計一 func() 名為 "text2Sentence(inputSTR)"，接受參數為字串，而 return 回傳值為一「斷句處理後」的 list。
    c. 設計一程式進入點，透過前述 "jsonTextReader() 讀取 example/news.json 檔中的 "text" 欄位的值，再透過 "text2Sentence()" 斷出完整的句子。
    d. 程式進入點的最後輸出要和 example/test.json 中的 "sentenceLIST" 欄位的值一致。