文本前處理：CKIPTagger 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

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:

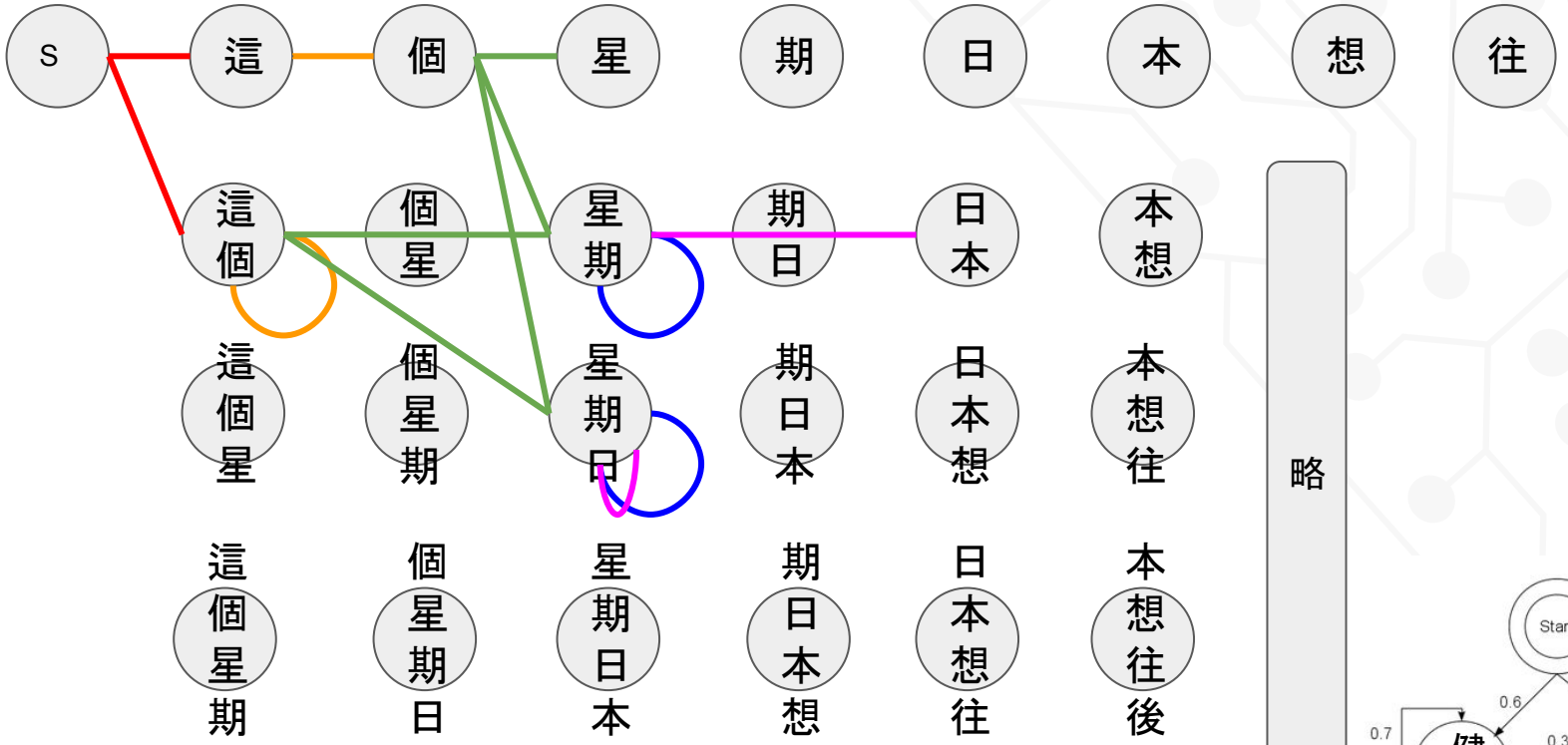
這/個/星/期/日/本/想/往/後/山/藥/師/佛/寺/去/世/人/罕/至/處/想/一/想/自/己/的/人/生

bi-gram:

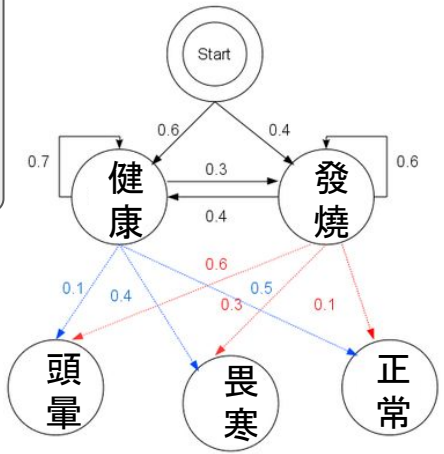
這個/個星/星期/期日/日本/本想/想往/往後/後山/山藥/藥師/師佛/佛寺/寺去/去世/世人/人罕/罕至/至處/處想/想一/一想/想自/自己/己的/的人/人生

tri-gram:

...



這/個/星...
這/個/星期/日本...
這個/星期/日本...
這/個/星期日/本...
這個/星期日/本...



BERT sentence pair encoding
(with tensors for PyTorch implementation)

CKIPTagger 結巴斷詞 (中研院詞庫小組)

https://github.com/ckiplab/ckiptagger

https://ckip.iis.sinica.edu.tw/service/corenlp/

- 安裝:
  - 感謝台師大資訊中心支援
- 使用:
  - 下載 Pietty (https://drive.google.com/open?id=0BxKoW6fgUa0CSTJDMmlDNC1nUDg)
  - 打開 Pietty，連線

IP 位置請見 moodle 公告
(呿…差一點就公開了 …)

系統操作指令

	功能	範例	解釋
cd	切換目錄 (change directory)	cd PeterOffice	切換到 PeterOffice 目錄
..	上一層目錄	cd ..	切換到「上一層」目錄
ls	列出內容	ls PeterOffice	列出 PeterOffice 的內容
python3 xxx.py	用 python3 執行 xxx.py 檔	python3 xxx.py	用 python3 執行 xxx.py 程式
exit	登出	exit	登出主機

nano 操作指令

	功能	範例	解釋
nano	啟用 nano 文字編輯器	nano xxx.py	用 nano 編輯 xxx.py 這個檔案
ctrl + o	寫入檔案 (存檔)		
ctrl + x	離開 nano		

利用「正規表示式」(regular expression) 擷取資訊

from "**data**" to "**information**"

這天章魚在街上撞到了一本書。

去台南參觀古蹟要不要錢啊？

搭到新竹火車站下車

實際操作

https://pythex.org/

(?<=XXX) 你會抓到的東西 (?=XXX)

實際操作

(?<=XXX) 你會抓到的東西 (?=XXX)

(?<=在).*?(?=上)

這天章魚在街上撞到了一本書。
這天章魚在桌面上看到了一本書。

(?<=[去來]).*?((?=參觀)|(?=喝))

去台南參觀古蹟要不要錢啊？
來台南喝飲料要不要加糖啊？

(?<=到).*?(?=[下轉]車)

搭到新竹火車站下車
坐到台北捷運站轉車

```
import re
inputLIST= ["這天章魚在街上撞到了一本書。",
"這天章魚在桌面上看到了一本書。"]

resultLIST = []
pat = re.compile("(?<=在).*?(?=上)")
for i in inputLIST:
    resultLIST.append([p.group(0) for p in pat.finditer(i)])

print(resultLIST)
```

實際操作

(?<=XXX) 你會抓到的東西 (?=XXX)

(?<=\(P\)).*?\(N[ac]\)(?= .\(Ncd\))

這(Nep) 天(Nf) 章魚(Na) 在(P) 街(Na) 上(Ncd) 撞到(VC) 了(Di) 一(Neu) 本(Nf) 書(Na) 。(PERIODCATEGORY)
這(Nep) 天(Nf) 章魚(Na) 在(P) 桌面(Nc) 上(Ncd) 看到(VE) 了(Di) 一(Neu) 本(Nf) 書(Na) 。(PERIODCATEGORY)

((?<=\(VCL\))|(?<=\(VA\))).*?(?=)

去(VCL) 台南(Nc) 參觀(VC) 古蹟(Na) 要不(Cbb) 要(VC) 錢(Na) 啊(T) ？(QUESTIONCATEGORY)
來(VA) 台南(Nc) 喝(VC) 飲料(Na) 要不要(D) 加(VC) 糖(Na) 啊(T) ？(QUESTIONCATEGORY)

((?<=\(VCL\))|(?<=\(V[AC]\))|(?<=\(Nc\))).*?(?=)

搭到(VC) 新竹(Nc) 火車站(Nc) 下車(VA)
坐到(VCL) 台北(Nc) 捷運站(Nc) 轉車(VA)

實際操作：抽出「地點」

這個星期日本想去藥師佛寺想一想自己的人生

這\個\星期\日本\想\去\藥師\佛寺\想\一\想\自己\的\人生

這(Nep) 個(Nf) 星期(Na) 日本(Nc) 想(VE) 去(VCL) 藥師(Na) 佛寺(Nc) 想(VE) 一(D) 想(VE) 自己(Nh) 的(DE) 人生(Na)

Coding session: More about regex (regular expression)

非中文的文字系統：用 re 描述目標

\w\d{7}

A1054101 薛○澤 林英杰
A1054102 張○君 林英杰
A1054103 李○儀 施簡信宏
A1054104 鍾○叡 張志鴻
A1054111 林○廷 張　蘭
A1054112 李○涵 鄭斯恩
A1054113 許○筠 陳晴玉
A1054114 陳○文 曾昱豪
A1054115 曾○宸 郭岳承
A1054147 鄭○琳 施信宏

中文的文字系統：用 re 描述前後文

(?<=[^\w]\s{5}).*?(?=\n)

A1054101 薛○澤 林英杰
A1054102 張○君 林英杰
A1054103 李○儀 施簡信宏
A1054104 鍾○叡 張志鴻
A1054111 林○廷 張　蘭
A1054112 李○涵 鄭斯恩
A1054113 許○筠 陳晴玉
A1054114 陳○文 曾昱豪
A1054115 曾○宸 郭岳承
A1054147 鄭○琳 施信宏

Quiz:

課堂中已經演示了文本處理中，利用 CKIPTagger 進行「斷詞工作」、POS 詞標標記以及 NER 命名實體辨識的幾個任務。試思考以下問題：

1. 斷詞系統的運作原理有哪幾種？CKIPTagger 是屬於哪一種？
2. 「訓練文本」和「應用文本」的差異，是否會造成 CKIPTagger 的表現不佳？
3. 只有斷詞處理，能取出什麼樣的資訊？這個資訊是否足以呈現文本特性？
4. 加上 POS 處理，能取出什麼樣的資訊？用什麼方法？
5. 再加上 NER 處理，能取出什麼樣的資訊？這個資訊是否有什麼限制？

https://github.com/ckiplab/ckiptagger/issues

https://github.com/ckiplab/ckiptagger/wiki/POS-Tags

Assignment: 小組作業，每人都要繳，但同組內容應一致

1. 從課程 github repo 中把課程中提供的 week08 的目錄 git pull 下來。
2. 把 week08.py 改名為 **week08_你的學號_分組隊名.py**
3. 在 **week08_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. 從伺服器裡，利用 CKIPTagger 處理你的隊伍名冊字串。
    b. 利用複製取出 POS 的結果字串，貼入你的 **week08_你的學號_分組隊名.py** 中
    c. 利用 Regular Expression 取出你的隊伍中的「人名」。
    d. 比較看看，你用 re 做出的結果和伺服器裡的 CKIPTagger 辨識出來的人名結果，何者較佳？
    e. 請設計一支名為 nameMail() 的函式，利用 re 模組取出你這一組的隊員的資料，回傳型態為一 list，內容為 ["姓名", "email"]
    f. 請設計一程式進入點，並給予 inputSTR = "<伺服器中，你這一組的 sentenceLIST 中的字串>" 為輸入值。
    g. 在程式進入點內將 inputSTR 傳入 nameMail() 中，取得結果存入 resultLIST 中以後，再印出 resultLIST.