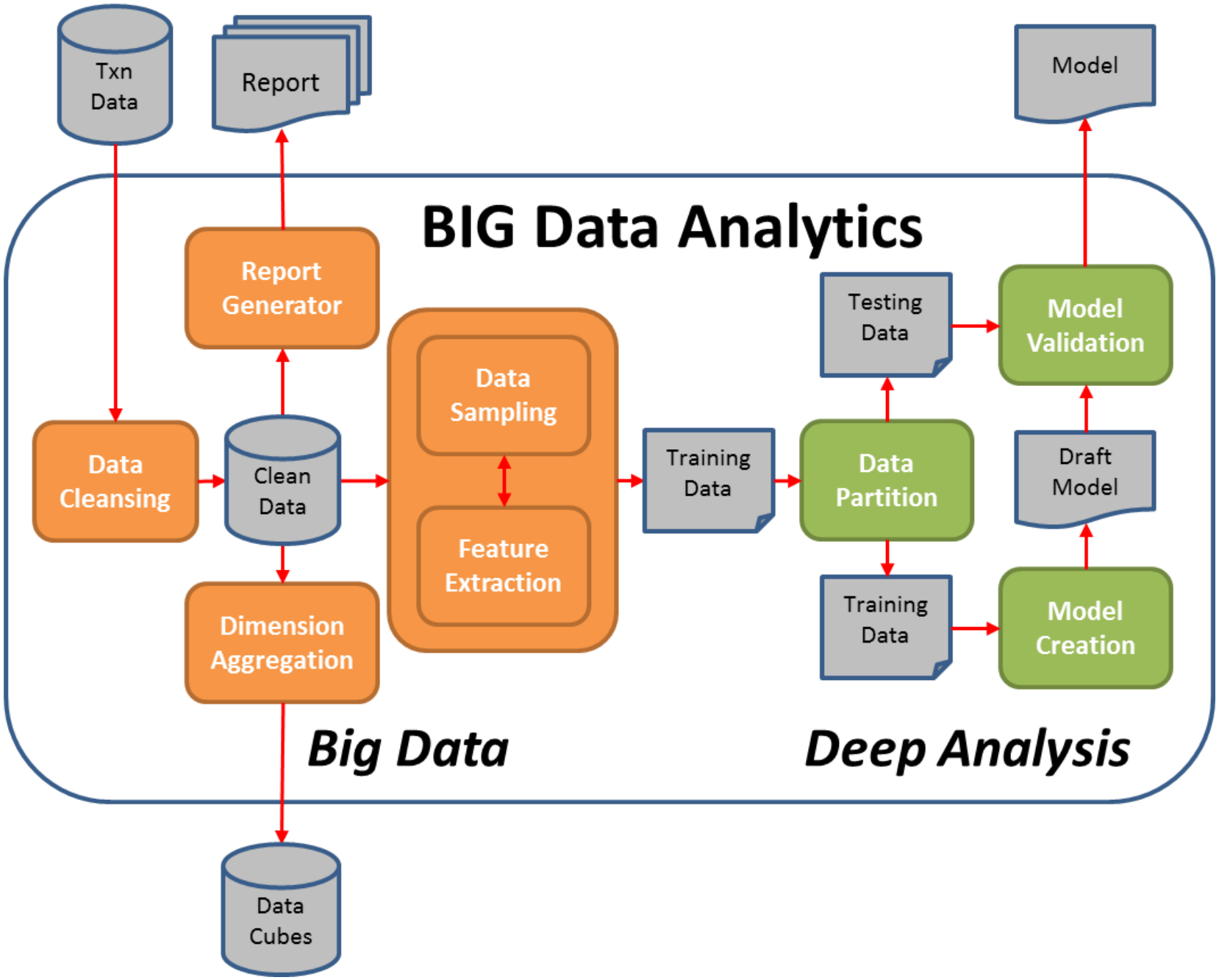
心理與神經資訊學
(Psychoinformatics & Neuroinformatics)

課號：Psy5261　　教室： 博雅 101

識別碼：227U9340　　時間： 四 234


再忍一下學期就結束了
~

資料的清理是使用的第一步



文字資料的處理
(RE & NLTK)

今天需要安裝的模組

若非使用 Anaconda Python， 請:

pip install re nltk

或

easy_install re nltk

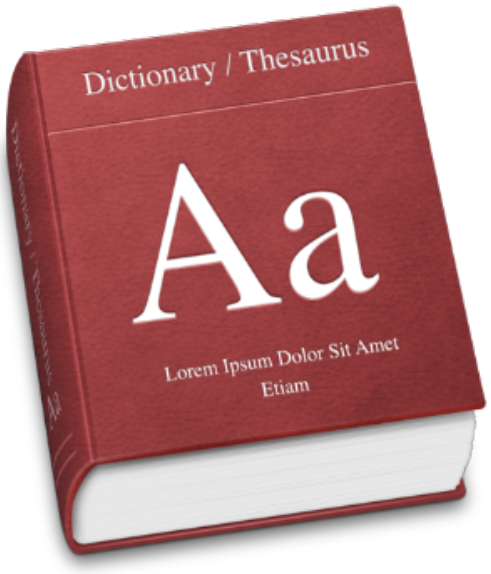
裝完後，請:

>>> import nltk

>>> nltk.download()

並下載 All packages

心理學案例研究



小美研究心理語言學。她想要:

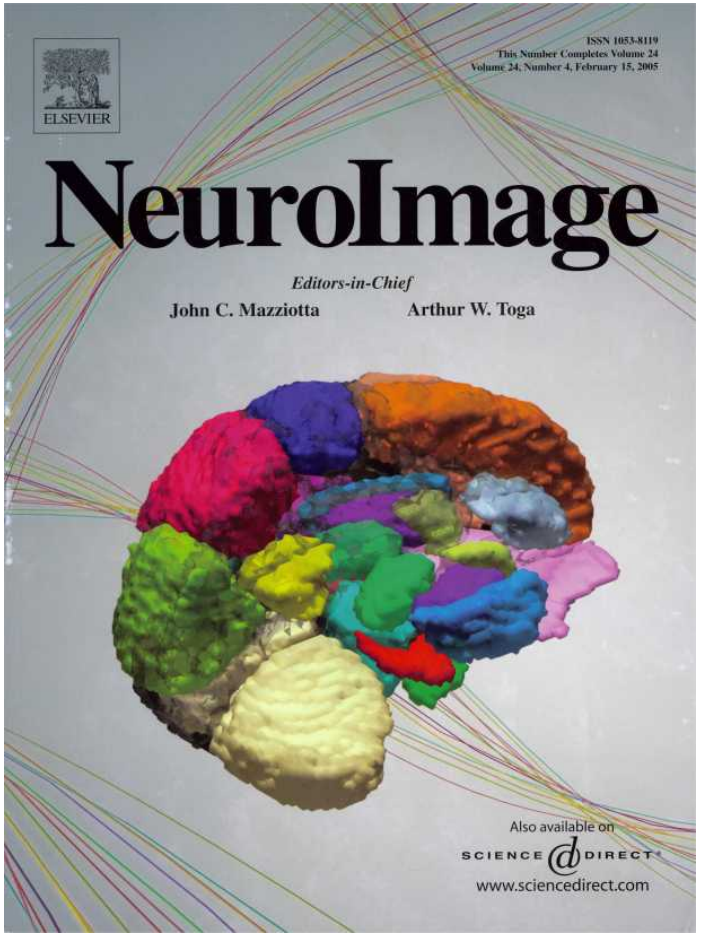
使用 3 個字元組成的英文字 (如 dog 、 cat 等)
及 3 個字元組成的非英文字 (如 frw 、 ukq 等)

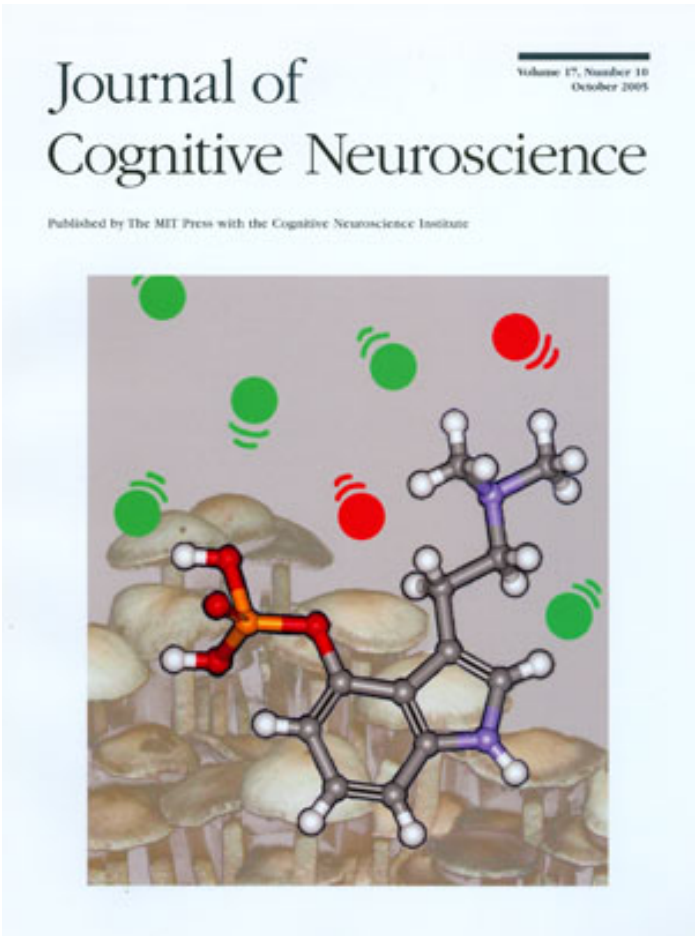
並找出單字辨認的反應時間和影響反應時間的因素

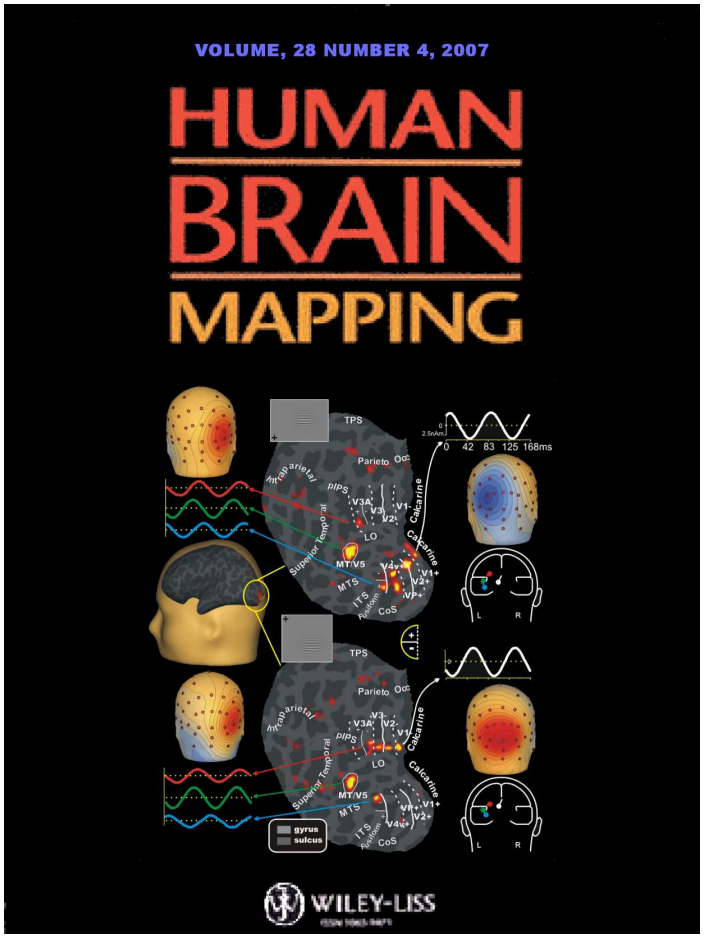
除了手動從字典找字外有沒有其他的方法?

腦科學案例研究

小明要寫程式讓電腦幫我們讀 papers







每家期刊格式都不同，怎麼辦？

Python String Module

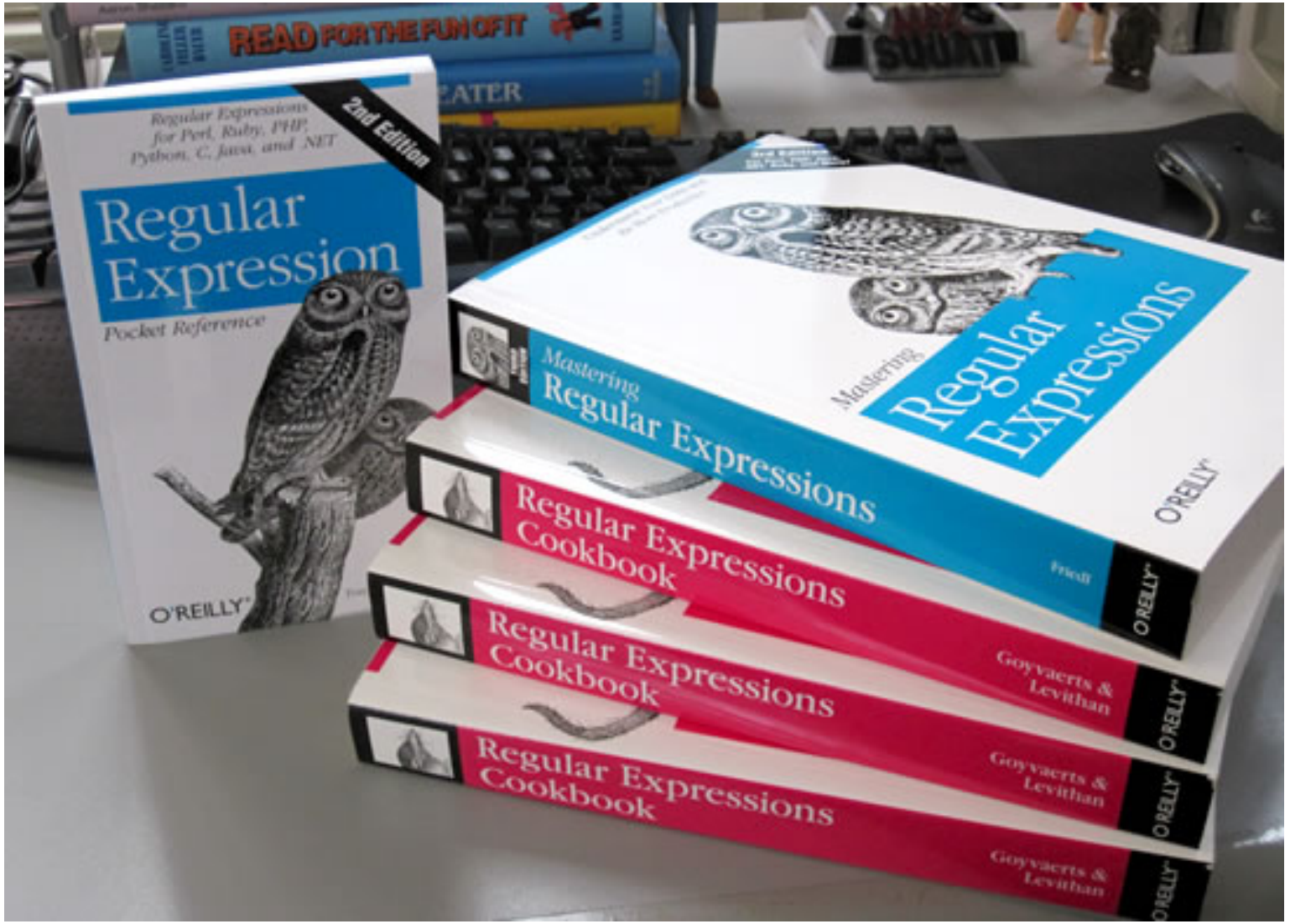
Python 內建就有一些文字資料處理能力

```
a='This is a built-in module'
print(a.lower(),a.upper())
print(a.split(' '),a.split('-'))
print(a.find('built'),a[10:15])
print(a.replace('module','library'))
print(a.count('is'))
print(a.center(50,'='))
```

```
7MM"""Mq.                                  `7MM                     db                 .d' ""                                                      mm      db
 MM   `MM.                                   MM                                        dM`                                                         MM
 MM   ,M9 ,pP"Ybd `7M'   `MF',p6"bo          MMpMMMb.   ,pW"Wq.  `7MM  `7MMpMMMb.     mMMmm   ,pW"Wq. `7Mb,od8 `7MMpMMMb.pMMMb.   ,6"Yb.  mmMMmm `7MM  ,p6"bo  ,pP"Ybd
 MMmmdM9  8I   `"   VA   ,V 6M'  OO          MM    MM  6W'   `Wb   MM    MM    MM      MM    6W'   `Wb  MM' "'  MM    MM    MM  8)   MM    MM     MM 6M'  OO  8I   `"
 MM       `YMMMa.    VA ,V  8M               MM    MM  8M     M8   MM    MM    MM      MM    8M     M8  MM      MM    MM    MM   ,pm9MM    MM     MM 8M       `YMMMa.
 MM       L.   I8     VVV   YM.    ,         MM    MM  YA.   ,A9   MM    MM    MM      MM    YA.   ,A9  MM      MM    MM    MM  8M   MM    MM     MM YM.    , L.   I8
JMML.     M9mmmP'     ,V     YMbmd'        .JMML  JMML. `Ybmd9'  .JMML..JMML  JMML..JMML.   `Ybmd9' .JMML.    .JMML  JMML  JMML.`Moo9^Yo.  `Mbmo .JMML. YMbmd'  M9mmmP'
                     ,V
                  OOb"
```

正規表示式 (Regular Expression)

RegEx 博大精深，今天只能簡介



正規表示式 (Regular Expression)

RegEx 的文字處理能力非常強大

Oh, no – the rumors were true. Our entire division's been replaced by a regular expression.

正規表示式 (Regular Expression)

看起來很像火星文

臺灣身份證號碼簡單檢查:

```
import re
print(bool(re.match('[A-Za-z]\d{9}','a123456789')))
```

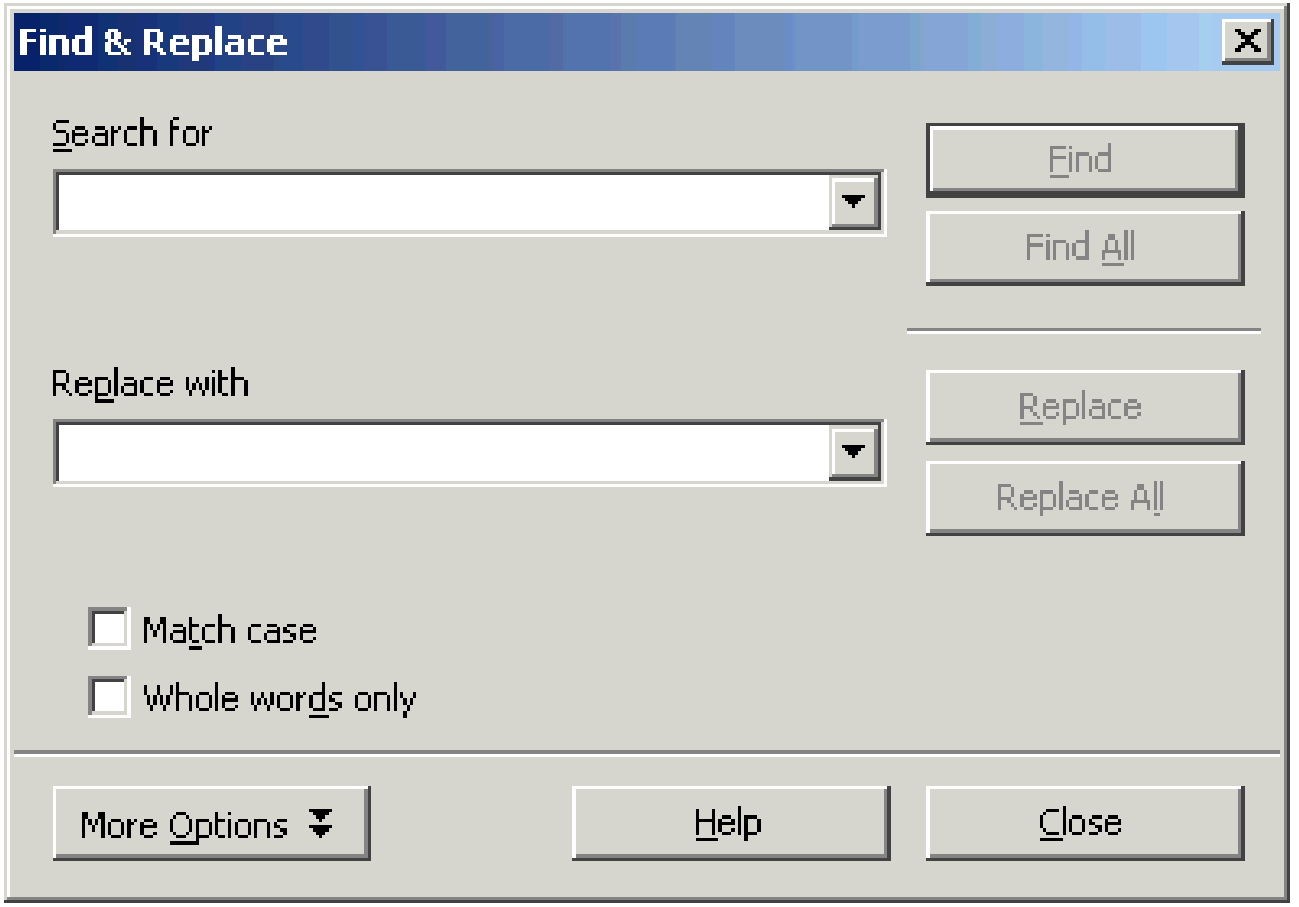
電子郵件地址的簡單檢查:

```
import re
print(bool(re.match('[^@]+@[^@]+\.[^@]+','a@b.c')))
# [^@]+  代表一個以上非 @ 的任何字元
```

Python 的 re module 文件看這裡

正規表示式 (Regular Expression)

功能可概略分成"搜尋"與"取代"



搜尋:

- re.compile()
- re.search()
- re.match()
- re.findall()
- re.split()
- ……

取代:

- re.sub()
- ……

強大處在於搜尋的是樣式 (pattern) 而非一特定字

用 RegEx 來搜尋

```
import re
regex=re.compile('abc',re.IGNORECASE)
for txt in ['abc','hello abc','hi AbC aBc']:
  print('-'*50)
  out=regex.search(txt) #inexact match
  if(out): print('search() found:',out.string)
  out=regex.match(txt) #exact match
  if(out): print('match() found:',out.string)
  out=regex.findall(txt) #search into a list
  if(out): print('findall() found:',out)
```

應用：搜尋 retweets（回推）

```
import re
tweets=['RT @spiketren  No class tomorrow',
'No class tomorrow (via @spiketren)']
rt=re.compile('(RT|via) (@\w+)')
#rt=re.compile('(RT|\(via) (@\w+\)*)')
for t in tweets:
  m=rt.search(t)
  print(m.group(1),m.group(2));
  #print(m[0],m[1],m[2]);
```

應用：身份證號碼檢查 / 產生器

```
import re,numpy as np
id='a123456789'
conv={'A':10,'B':11,'C':12,'D':13,'E':14,'F':15,'G':16,
'H':17,'I':34,'J':18,'K':19,'L':20,'M':21,'N':22,'O':35,
'P':23,'Q':24,'R':25,'S':26,'T':27,'U':28,'V':29,'W':32,
'X':30,'Y':31,'Z':33}
parts=re.findall('\w',id)
region=str(conv[parts[0].upper()])
coef=np.array([1]+list(range(9,0,-1))+[1])
digits=np.array(list(region)+parts[1:],dtype=np.int)
if(np.mod(np.sum(coef*digits),10)==0): print('Valid!')
```

用 RegEx 來取代

修完本課失業後至少可以當 spammer?

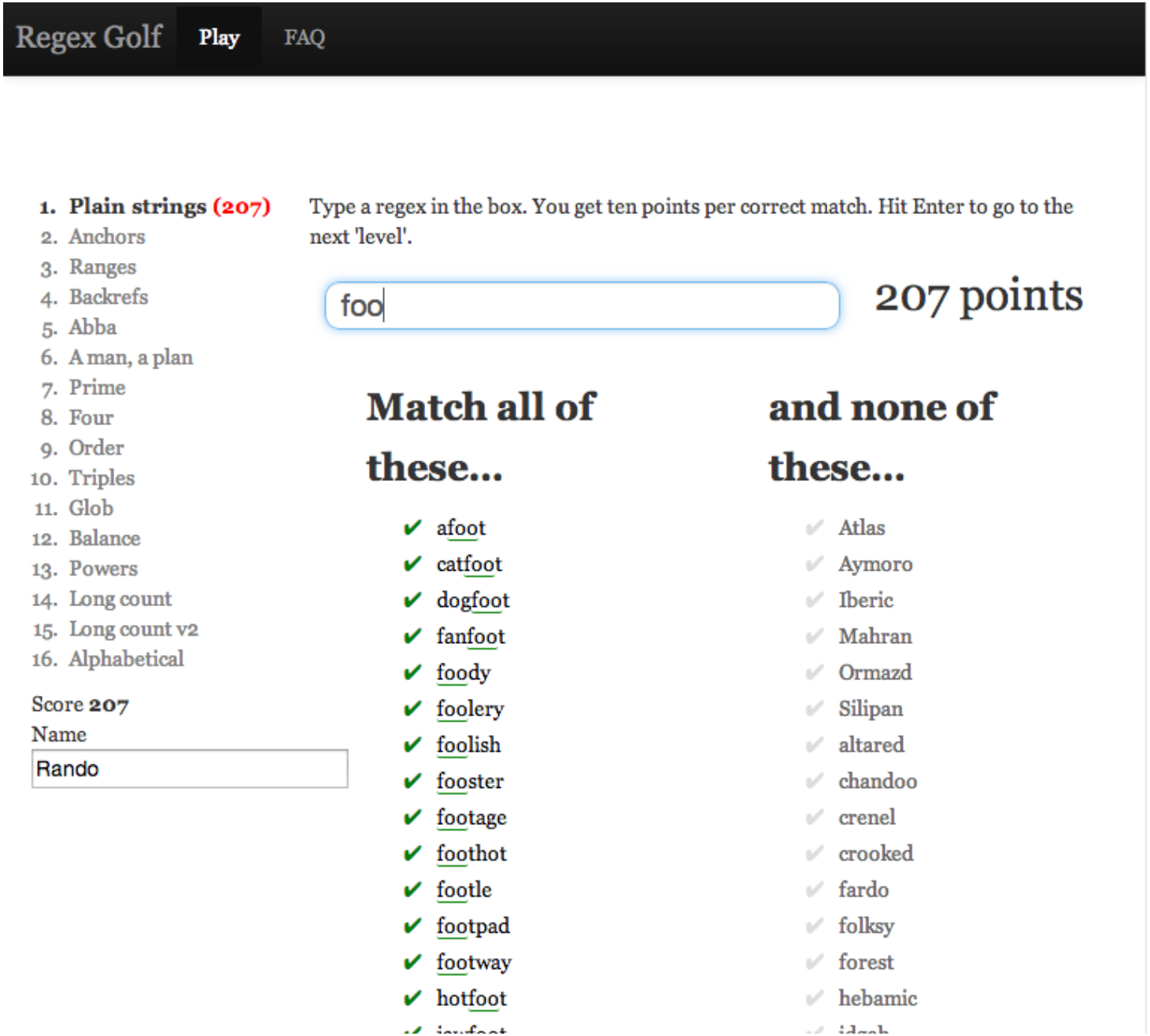
```
import re
html='<body><b>test</b><img src=test.jpg></body>'
print(re.sub('<[^<]*>',"",html))
e='@mail.ncku.edu.tw'
t='pichun_huang,chendy'
print(re.sub('\w+','\g<0>'+e,t))
```

職稱	助理教授
姓名	黃碧群
e-mail	pichun_huang
專長	視知覺，弱視，空間視覺的

職稱	助理教授
姓名	陳德祐
e-mail	chendy
專長	人類及動物的功能性腦造影

*為避免垃圾郵件濫發，教師 E-mail 請自行加上 @mail.ncku.edu.tw

RegEx 益智遊戲


Regex Golf
Play
FAQ
1. Plain strings (207)
2. Anchors
3. Ranges
4. Backrefs
5. Abba
6. A man, a plan
7. Prime
8. Four
9. Order
10. Triples
11. Glob
12. Balance
13. Powers
14. Long count
15. Long count v2
16. Alphabetical
Score 207
Name
Rando
Type a regex in the box. You get ten points per correct match. Hit Enter to go to the next 'level'.
foo
207 points
Match all of these...
afoot
catfoot
dogfoot
fanfoot
foody
foolery
foolish
fooster
footage
foothot
footle
footpad
footway
hotfoot
and none of these...
Atlas
Aymoro
Iberic
Mahran
Ormazd
Silipan
altared
chandoo
crenel
crooked
fardo
folksy
forest
hebamic

自然語言處理(NLP)

自然語言處理（Natural Language Processing）
是人工智慧和語言學領域的分支學科。
在這此領域中探討如何處理及運用自然語言；
自然語言認知則是指讓電腦「懂」人類的語言。

句子「我們把香蕉給猴子，因為它餓了」和
「我們把香蕉給猴子，因為它熟透了」
有同樣的結構。但是代詞「它」在第一句中指的是
「猴子」，在第二句中指的是「香蕉」。
如果不了解猴子和香蕉的屬性，無法區分。

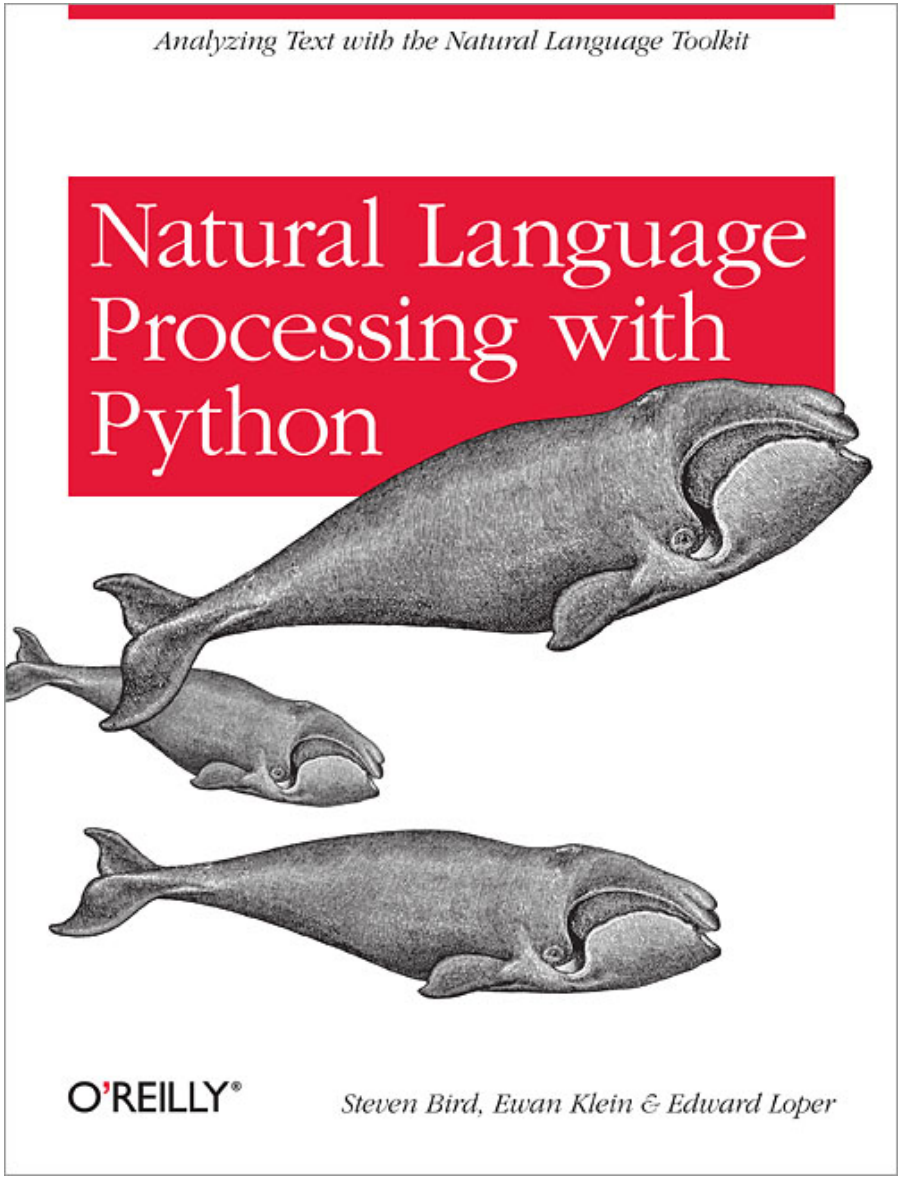
自然語言處理的應用


中華電信
下午5:21
有什麼我可以幫忙的嗎？
“世界上有外星人嗎？”
關於這個，寶杰，您的看法是…？
“世界上有外星人嗎？”
這可能需要問一下輒青，您怎麼看？


中華電信
下午5:27
好的…不過我早就叫你“主人”了，主人。
“妳會說台灣國語嗎”
台灣國語？我當然會呀，這樣才是愛台灣。

使用 Python 做自然語言處理

NLTK=Natural Language Toolkit



教育程度與意識形態

VS

聽說教育程度愈高意識形態愈嚴重

可是教育程度如何衡量呢？

```
from nltk.book import text4
print(len(set(text4))/len(text4)) # lexical diversity
long_words=[w for w in set(text4) if len(w)>15]
print(len(long_words)) # number of big words
```

斷句 / 詞 (Tokenization) 與統計

```
import nltk
mytxt='This is a cat. That is a dog.'
print(nltk.sent_tokenize(mytxt))
text=nltk.Text(nltk.word_tokenize(mytxt))
text.plot()
from nltk.book import text4
text4.dispersion_plot(['democracy','freedom','duties'])
dist=nltk.FreqDist(text4)
print(dist['freedom'])
print([w for w in dist.keys() if dist[w]>1000])
```

有加權的詞頻 :TF-IDF

term frequency–inverse document frequency

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

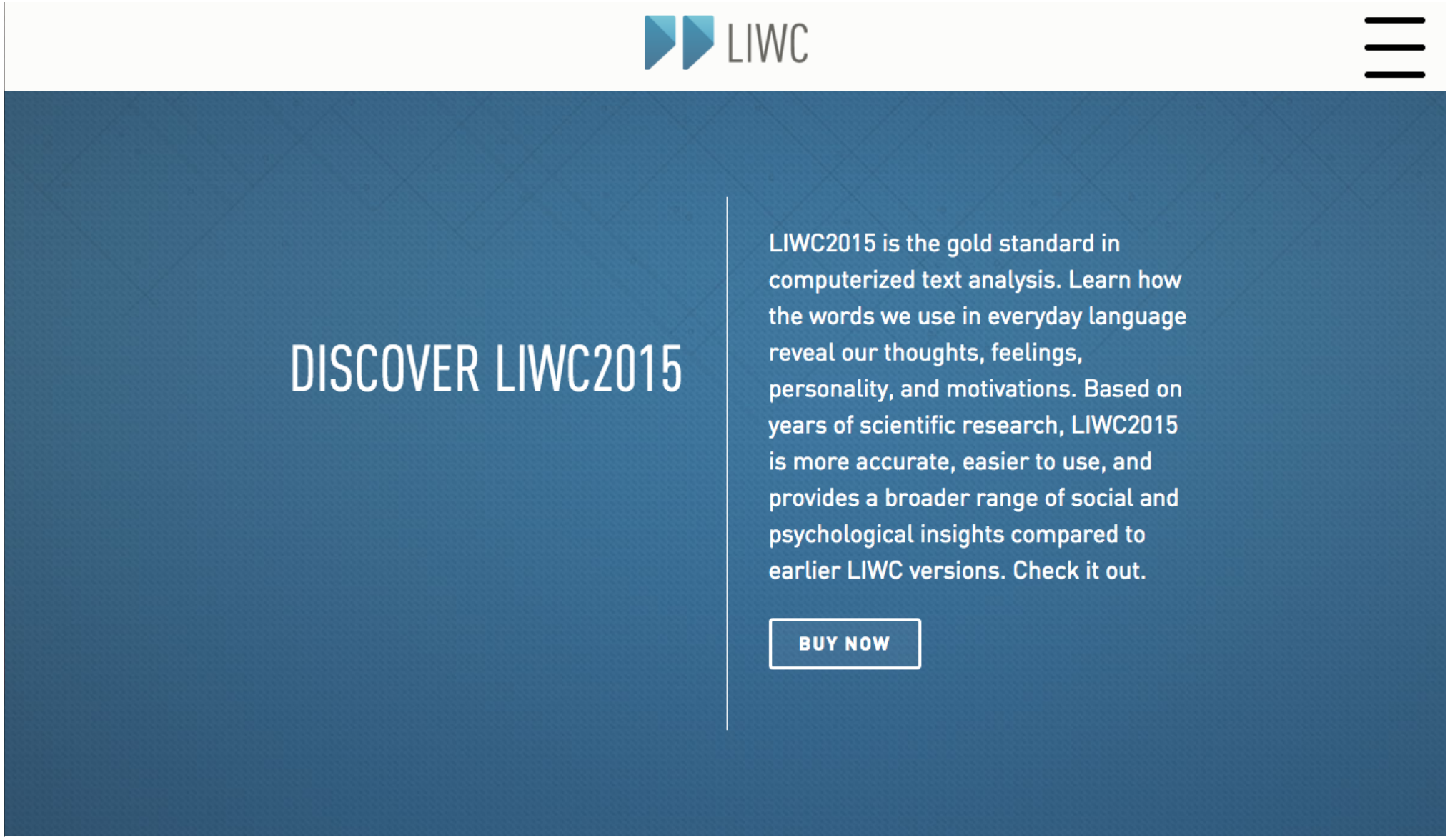
df_x = number of documents containing x

N = total number of documents

frequency	highest idf	tf * idf
32 the	1.00 tdt000077	3.20 orr
16 were	1.00 picknickers	2.81 charlotte
14 said	0.93 sreaming	2.65 payne
12 and	0.93 timmy	2.48 dc
12 to	0.86 6thld	2.24 usair
11 a	0.80 orr	2.00 plane
10 of	0.78 1016	1.93 crash
9 at	0.76 bergen	1.74 bones
9 was	0.75 dripping	1.63 survivors
7 in	0.73 abrams	1.50 dripping
6 on	0.72 0419	1.49 wreckage
6 they	0.69 fuselage	1.35 dead
6 people	0.66 nc	1.29 hospitals
6 had	0.66 thunderstorm	1.27 airport
6 plane	0.66 payne	1.23 55

Linguistic Inquiry & Word Count

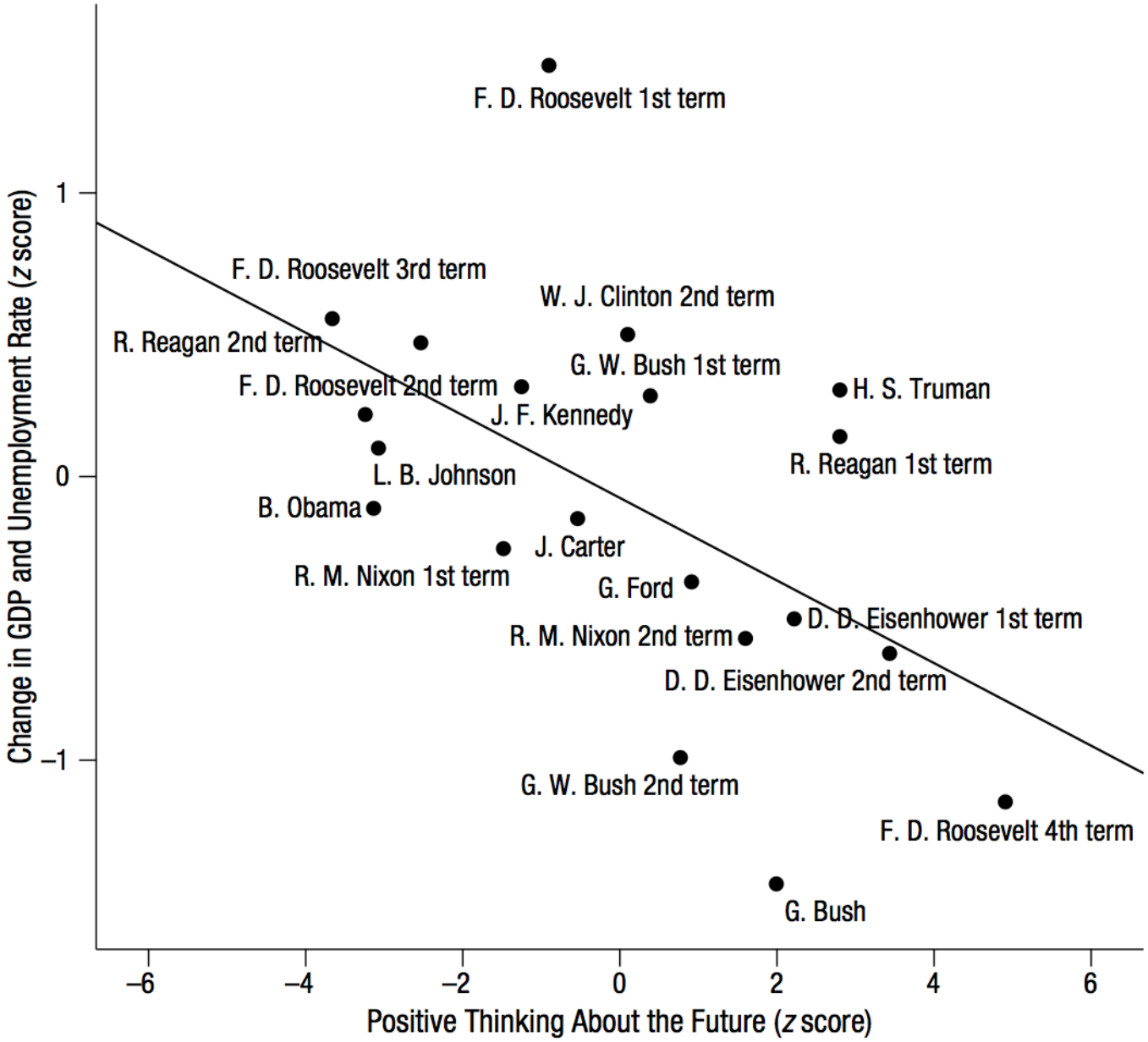
LIWC 有線上試用版



中文資料的處理：第一步要斷詞

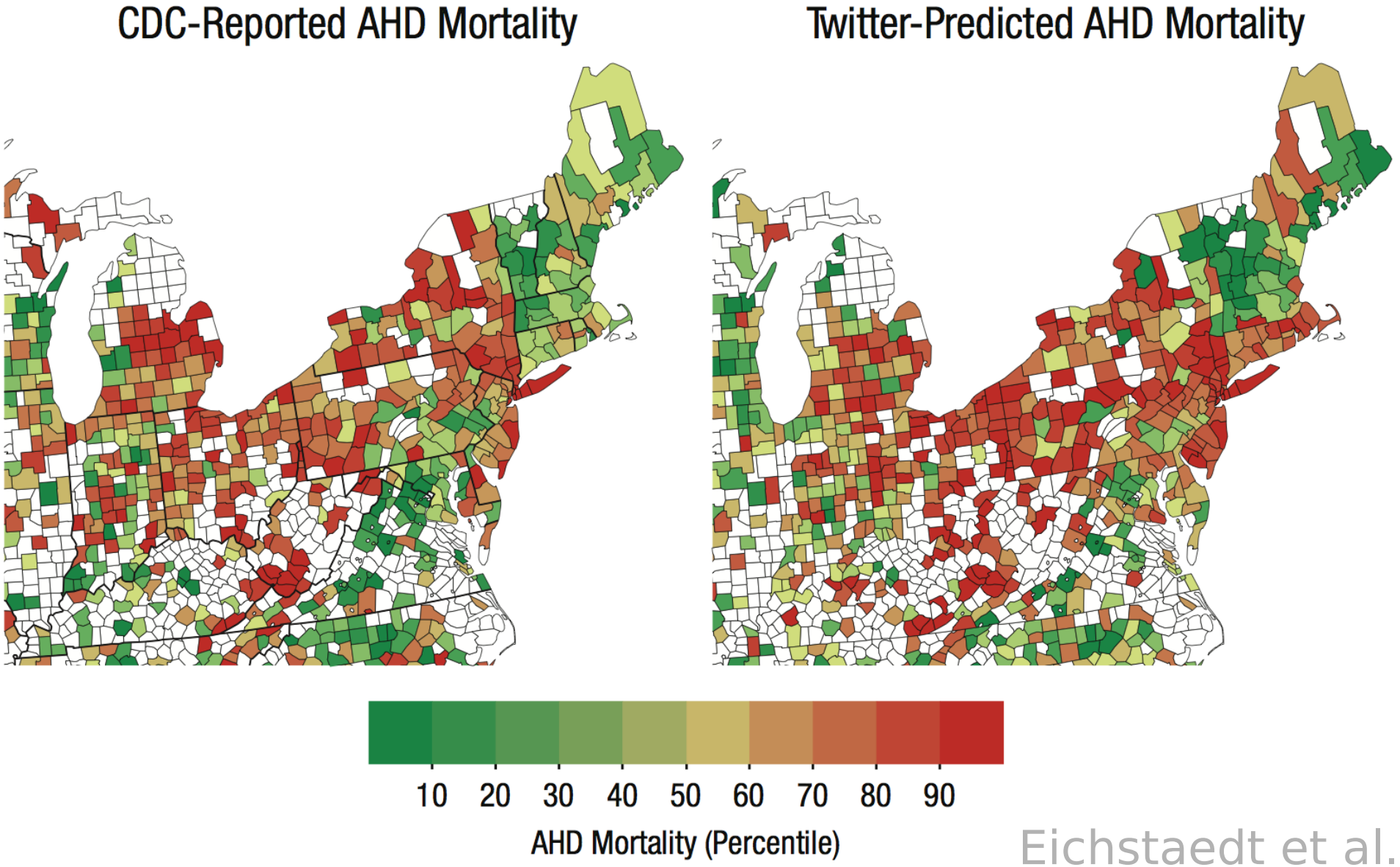
- [中研院中文斷詞系統]
- [結巴中文分詞套件]
- [中文版LIWC詞典（demo）]

心理學案例研究：未來正向思考 & GDP



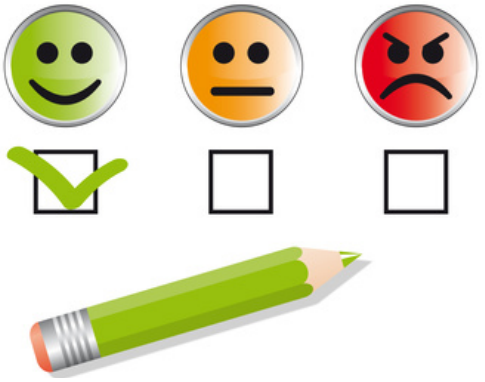
Sevincer et al., 2015, *Psychol. Sci.*

心理學案例研究：負向情緒 & 心臟病



Eichstaedt et al., 2015, *Psychol. Sci.*

情緒分析 (Sentiment Analysis)



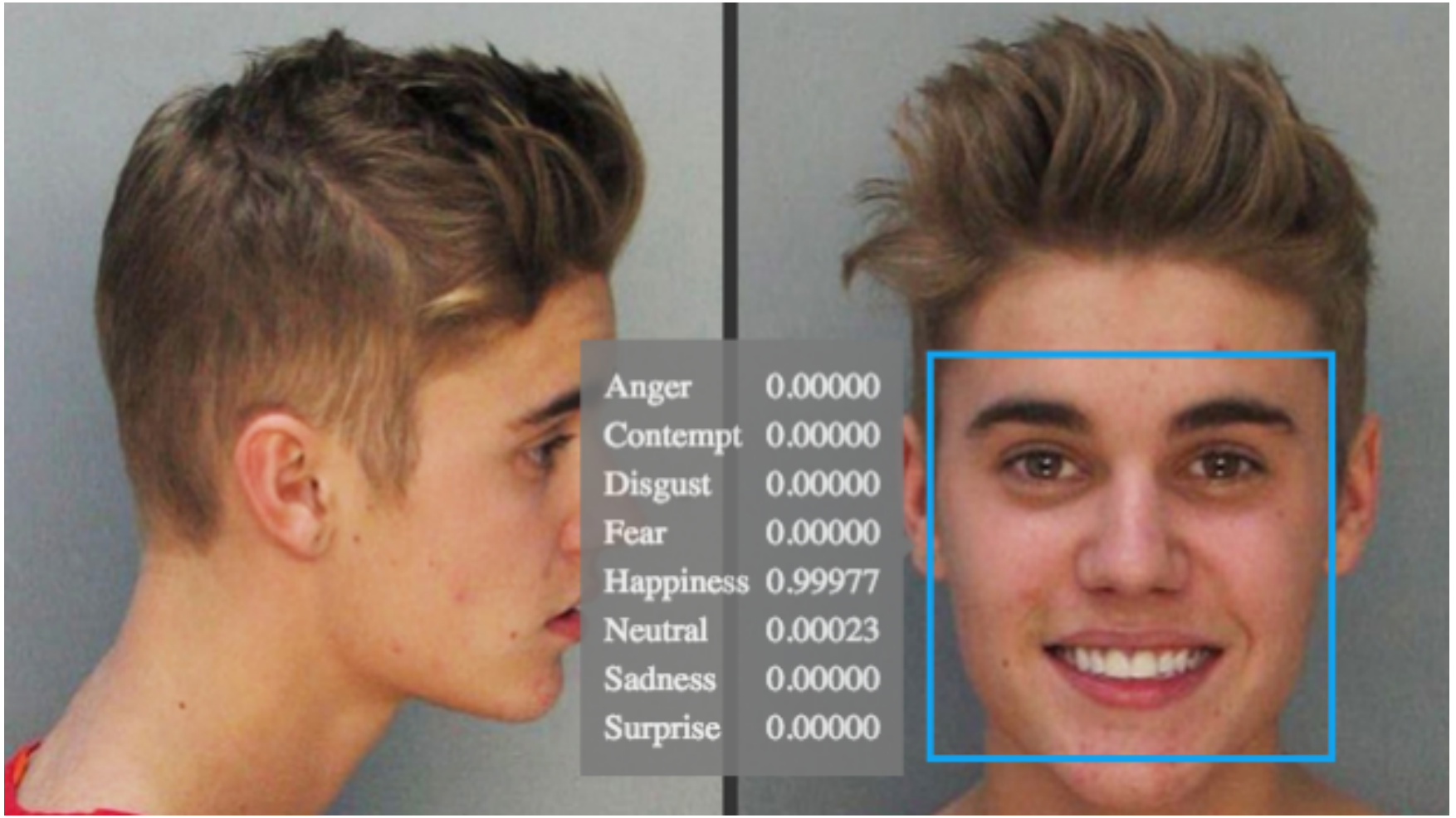
解法 2： 網路上先下載正面和負面情緒的字詞列表。
對一篇文章算出正面字和負面字各有多少，
若正面字數遠大於負面字數文章即為正面。

解法 3： 利用 nltk 中的機器學習 (nltk.classify) 來
自動分類句子或文章為正面或負面，例如
這個範例。

不只是正向或負向情緒

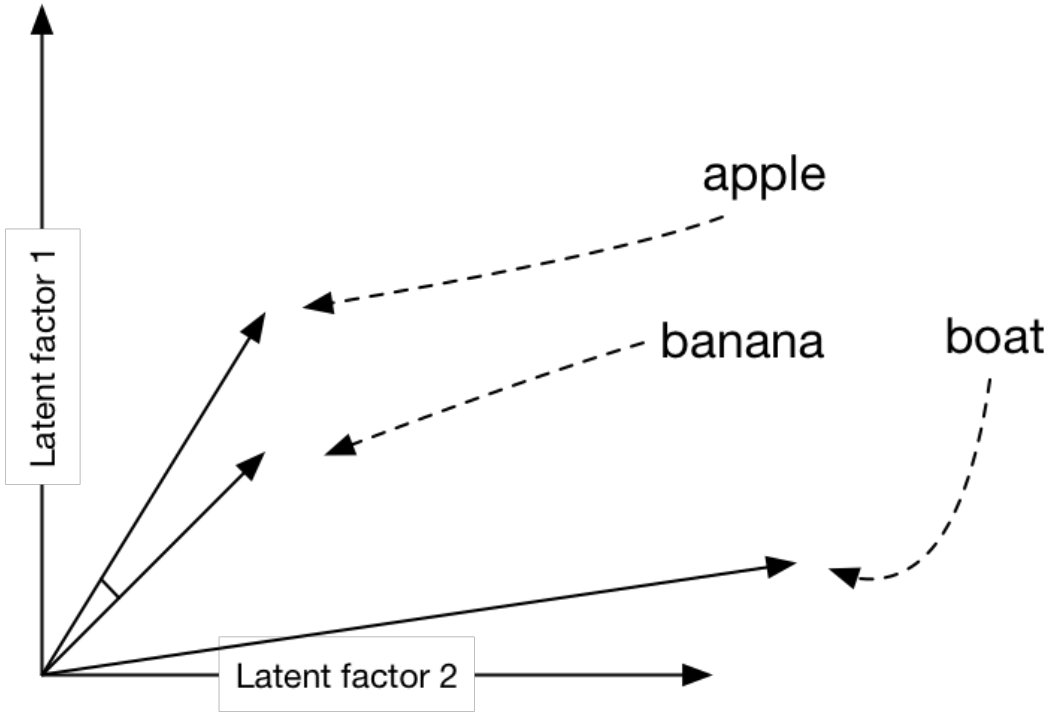
NRCWE將 14,182 個字分成 Plutchik’s 8 個情緒類別

Stanford U 也有做多情緒分析器



Semantic Similarity

Word2Vec的 Python 界面為 gensim; 應用請參考這

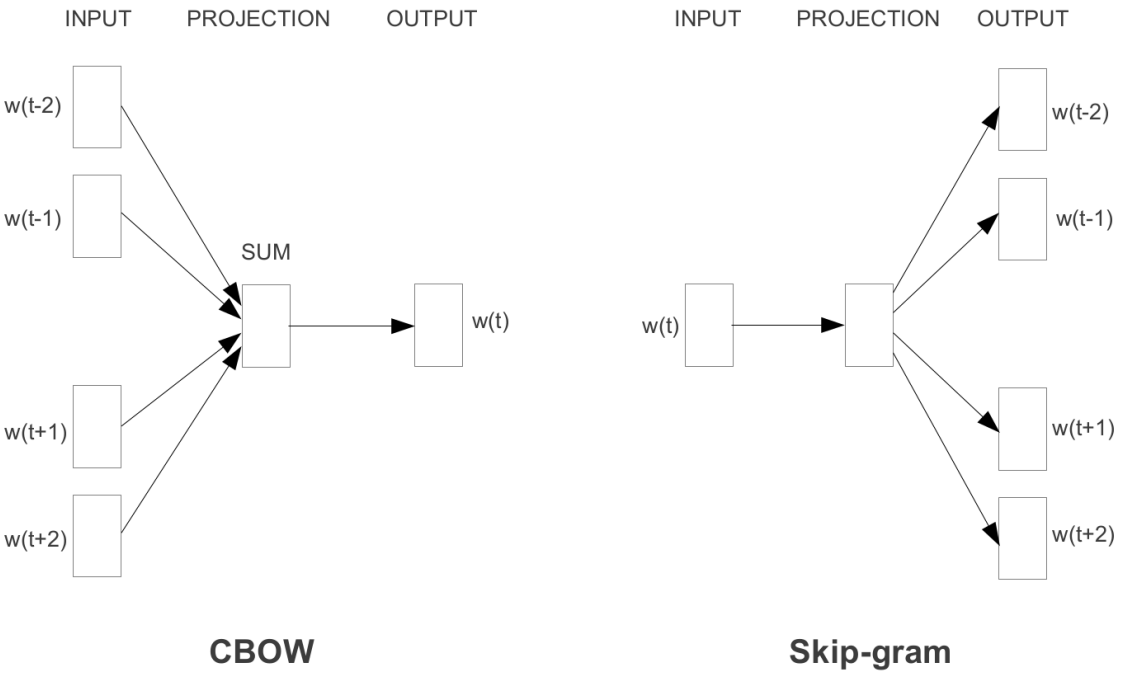


"king" - "man"+"woman" = "queen"

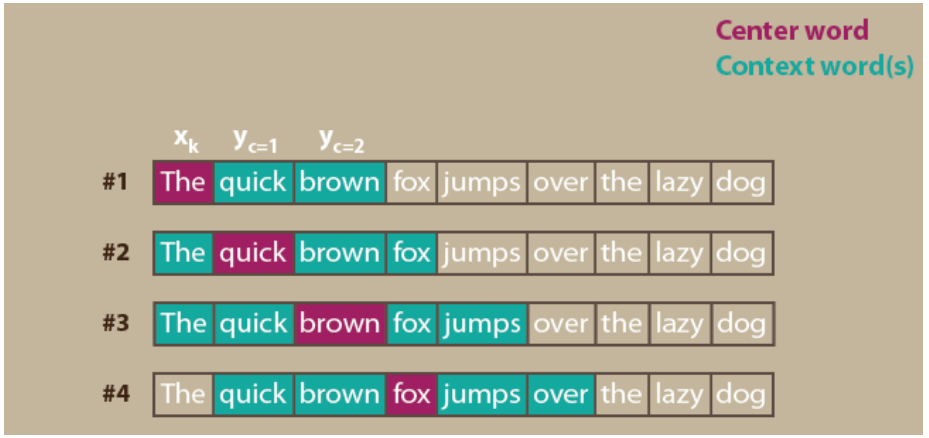
("good" + "best") / 2 = "better"

Word2Vec

得到 word vectors/representations 有兩種方式

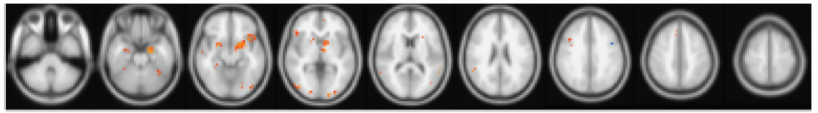


CBOW **Skip-gram**

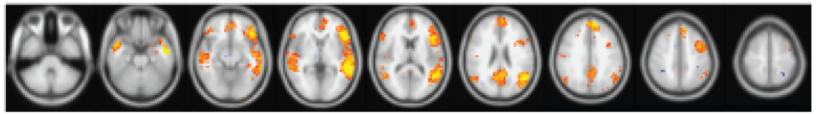


Topic Modeling (LDA)

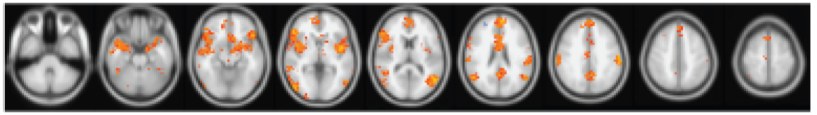
Topic 121 (68 docs)
regret, surprise, reasoning
arousal, learning



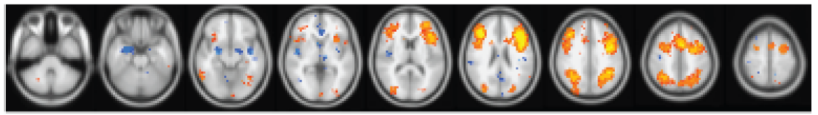
Topic 71 (115 docs):
narrative, discourse, comprehension,
memory, discourse_processing



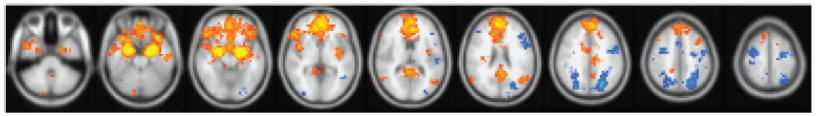
Topic 108 (128 docs):
empathy, pain, theory_of_mind,
awareness, facial_expression



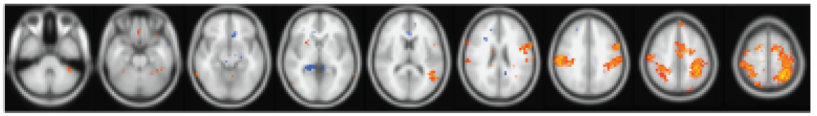
Topic 61 (441 docs):
memory, working_memory, maintenance,
visual_working_memory,
spatial_working_memory



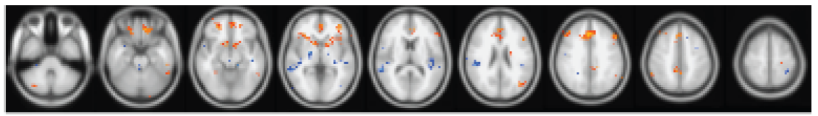
Topic 93 (495 docs):
emotion, valence, arousal, attention,
focus



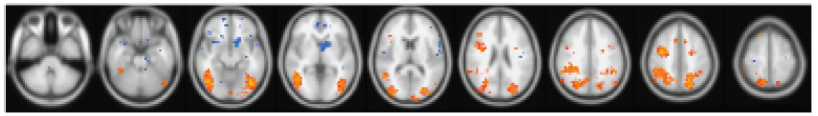
Topic 20 (497 docs):
action, movement, goal, context,
perception



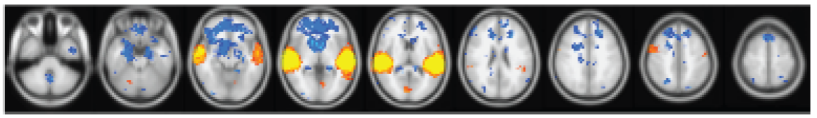
Topic 86 (519 docs):
decision, decision_making, choice,
fixation, uncertainty



Topic 43 (566 docs):
attention, focus, visual_attention,
fixation, attentional_resources



Topic 74 (769 docs):
auditory, perception, hearing,
attention, listening

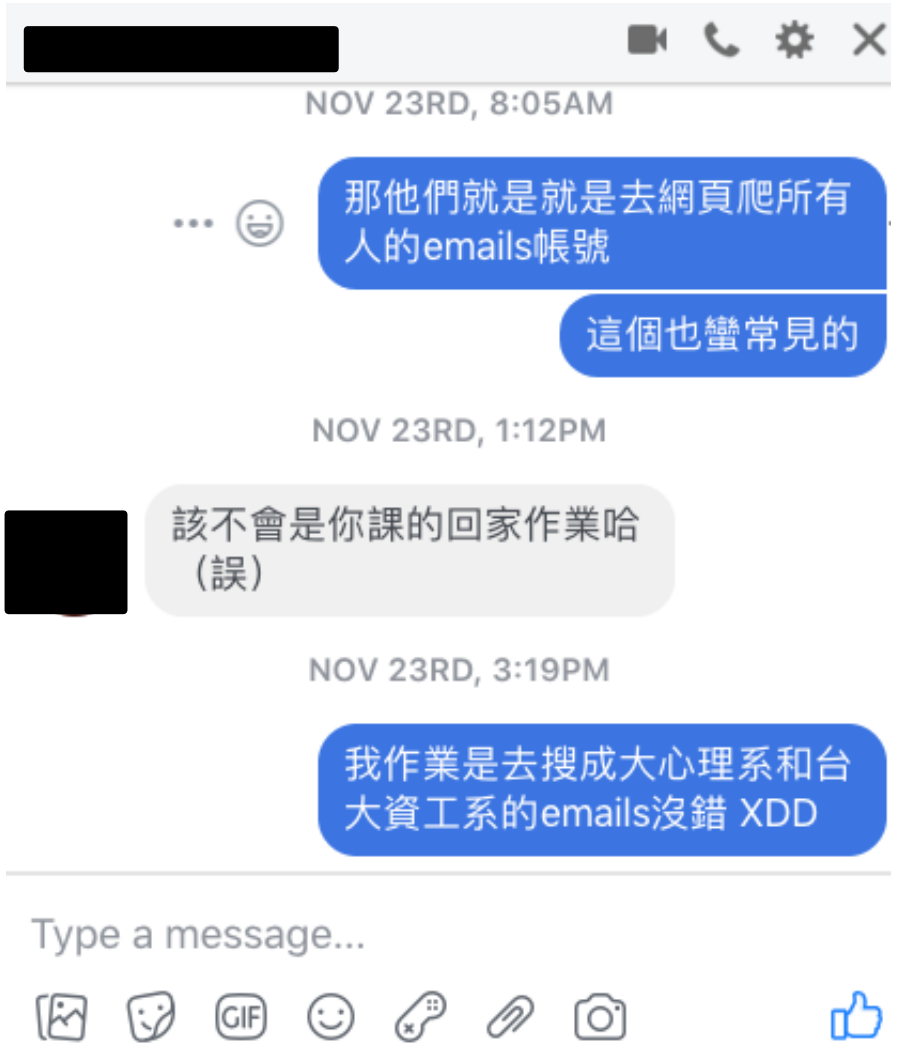


本週作業

進一步研究 regrex 與 topic modeling

1. 用 urllib+re 在 臺大資工系網頁搜集教師 emails

2. 對 2.6 Topic Modeling 的範例程式文本先過濾掉 stop words，使 5 個主題能夠更加明確



Game Over