自然語言處理
簡介 II

Unix for Poets
主修文科也需要的指令型工具

(original by Ken Church, modifications by Chris Manning/
Dan Jurafsky, Stanford University

Unix for Poets
(根據 Ken Church 的一份廣為流傳在美國各大學的講義)

- 到處都是文字
  - 網路
  - 辭典、語料庫、電子郵件等
  - 上億詞不稀奇
- 如何運用這些資料？
- 與其束手無策，不如做一些最簡單的事情
- 用 Unix 一行一行的指令可以做一些有用的事情
- 有時候，下指令比寫 Python 程式，來得快又有效
- 動手做，很好玩

接下來，我們示範如何動手做

1. 數一數「文本」中有哪些詞彙、詞彙的頻率
2. 排序詞彙表
   - 字母順序 ascii order 音韻順序 rhyming order
3. 擷取辭典中的相關資訊
4. 計算n連詞的統計資訊 ngram statistics
5. 運用有標註文本 tagged text 內的詞性資料

Unix 的指令工具

- **grep**: search for a pattern (regular expression)
- **sort**
- **uniq –c** (count duplicates)
- **tr** (translate characters)
- **wc** (word – or line – count)
- **sed** (edit string -- replacement)
- **cat** (send file(s) in stream)
- **echo** (send text in stream)
- **cut** (columns in tab-separated files)
- **paste** (paste columns)
- **head**
- **tail**
- **rev** (reverse lines)
- **comm**
- **join**
- **shuf** (shuffle lines of text)

準備

- 電腦環境
  - 蘋果電腦的終端機（iTerm）、微軟視窗——安裝 cwin、
  - 伺服器——請助教提供帳號
- 資料 NYT_19980407.txt （見課程網站）
- 在 Unix 指令中指定輸入、輸出 Input/output redirection
  - 輸出 >；輸入 < ；管線 | （在一個指令輸出後，接到下一個指令作為輸出
  - 用 echo/cat 輸入 (資料在指令/檔案中)
- 操作錯誤、操作太久，中斷指令的執行
  - CTRL-C

示範練習 1: Count words in a text

- 輸入: text file (nyt_200811.txt) (after it's gunzipped)
- 輸出: list of words in the file with freq counts
- 用指令來寫演算法
  1. 分詞 Tokenize (`tr`)
  2. 排序 Sort (`sort`)
  3. 數重覆次數 (`uniq –c`)

解答 1a

- `cat NYT_19980407.txt | tr -sc 'A-Za-z' '\n' | sort | uniq -c | head`

```
      1
     19 A
      1 AGAIN
      1 ALL
      3 AMP
      1 AND
      1 ANGELES
     10 ANNOTATION
      1 ANONYMOUS
      1 ARTISANS
```

解答 1b

- `cat NYT_19980407.txt | tr -sc 'A-Za-z' '\n' | sort -f | uniq -c | head`

```
   1
  19 A
 338 a
   1 abandoned
   1 abandoning
   1 Abby
   1 Abe
   8 able
   2 About
  35 about
```

解答 1c

- `cat NYT_19980407.txt | tr -sc 'A-Za-z' '\n' | sort | uniq -c | sort -nr | head -30`

```
1274 enamex
 960 type
 960 e
 958 b
 663 the
 406 of
 381 PERSON
 338 numex
 338 a
 316 and
 304 timex
 284 to
 239 in
 139 ORGANIZATION
 137 is
 133 s
 119 DATE
 117 LOCATION
 111 that
 102 on
  93 was
  90 for
  90 CARDINAL
  84 with
  81 be
  78 The
  75 by
  75 as
  73 I
  72 it
```

延伸練習

1. 合併大寫、小寫字（把大寫字母改成小寫）
   - 提示: 用第二個 tr 指令

2. 連續母音（如 ieu）的頻率
   - 提示: 用第三個 tr 指令

不同的排序

- `sort`
- `sort –f` 不論大小寫（由 a/A 到 z/Z）
- `sort –n` 視為數字，按數字大寫排列（由低到高）
- `sort –r` 反順序（由 z 到 a）
- `sort –nr` 反數字順序（由高到低）
- rev 把每一行的文字倒過來
  - `echo 'Hello' | rev`

Counting and sorting exercises

- Find the 50 most common words in the NYT
  - Hint: Use sort a second time, then head

- Find the words in the NYT that end in “zz”
  - Hint: Look at the end of a list of reversed words

心得

- 把指令用管線（pipe）連接起來，很簡單但又很厲害
- 組合指令很有彈性（寫一行的程式）
- Unix 哲學
  - 可以組合的小工具
  - 小兵立大功

(連續) 雙字統計表 Bigrams

演算法

1. 切分詞彙
2. 做成兩份（相差一行的兩個檔案）——使用 **tail** 指令
3. 把兩檔案接起來使得 $word_i$ 和 $word_{i+1}$ 在一行——用 **paste** 指令
4. 數一數——使用 **sort | uniq -c** 指令

指令與執行結果

- `tr -sc 'A-Za-z' '\n' < nyt_200811.txt > nyt.words`
- `tail -n +2 nyt.words > nyt.nextwords`
- `paste nyt.words nyt.nextwords > nyt.bigrams`
- `head –n 5 nyt.bigrams`

```
KBR     said
said    Friday
Friday  the
the     global
global  economic
```

練習

- 找到最常見的 10 個 2-連詞 bigrams
  - 然後用肉眼觀察 10 個 2-連詞的詞性有什麼規律
- 找到最常見的 10 個 3-連詞 trigrams

搜尋比對 grep

- 用正規式 regular expressions 搜尋檔案的一行一行（基本上一次比一行）

grep rebuilt nyt_200811.txt
Conn and Johnson, has been **rebuilt**, among the first of the 222
move into their **rebuilt** home, sleeping under the same roof for the
the part of town that was wiped away and is being **rebuilt**. That is
to laser trace what was there and **rebuilt** it with accuracy," she
home - is expected to be **rebuilt** by spring. Braasch promises that a
the anonymous places where the country will have to be **rebuilt**,
"The party will not be **rebuilt** without moderates being a part of

grep

- Grep 的名稱式一個縮寫
  - globally search for regular expression and print

- 搜尋、顯示 –ing 結束的詞
- `grep 'ing$' nyt.words |sort | uniq –c`

示範 grep 的使用

- `grep 'ing$' nyt.words | sort | uniq -c | sort -nr | head`

```
 924 going
 606 being
 578 including
 519 during
 395 according
 389 making
 374 something
 350 running
 318 voting
 293 trying
```

用 grep 來過濾輸入檔產生新檔

- `grep gh` 留下有 gh 的那幾行
- `grep '^con'` 留下 con 的開頭的那幾行
- `grep 'ing$'` 留下 ing 的結束的那幾行
- `grep -v gh` 留下沒有 gh 的那幾行
- `grep -e` 用正規式比對
- `grep -e '^[A-Z]+$' nyt.words |sort|uniq -c`
  留下大寫字開始的那幾行

Regular Expressions: Disjunctions

- Letters inside square brackets []

Pattern	Matches
`[wW]oodchuck`	Woodchuck, woodchuck
`[1234567890]`	Any digit

- Ranges `[A-Z]`

Pattern	Matches	
`[A-Z]`	An upper case letter	`Drenched Blossoms`
`[a-z]`	A lower case letter	`my beans were impatient`
`[0-9]`	A single digit	`Chapter 1: Down the Rabbit Hole`

Regular Expressions: Negation in Disjunction

- Negations `[^Ss]`
  - Carat means negation only when first in []

Pattern	Matches	
`[^A-Z]`	Not an upper case letter	`Oyfn pripetchik`
`[^Ss]`	Neither 'S' nor 's'	`I have no exquisite reason"`
`[^e^]`	Neither e nor ^	`Look here`
`a^b`	The pattern a carat b	`Look up a^b now`

Regular Expressions: More Disjunction

- Woodchucks is another name for groundhog!
- The symbol | (pipe or vbar) for disjunction (OR)

Pattern	Matches
`groundhog|woodchuck`	
`yours|mine`	yours mine
`a|b|c`	= [abc]
`[gG]roundhog|[Ww]oodchuck`	

Kleene invented Regular Expressions: ? * + .

Pattern	Matches	
colou?r	Optional previous char	color colour
oo*h!	0 or more of previous char	oh! ooh! oooh! ooooh!
o+h!	1 or more of previous char	oh! ooh! oooh! ooooh!
baa+		baa baaa baaaa baaaaa
beg.n		begin begun begun beg3n

Stephen C Kleene: any **regular** language is accepted by an FA and vise versa.

Kleene *, Kleene +

計算檔案的行數、詞數、字母數 lines, words, characters

- `wc nyt_200811.txt`

```
140000 1007597 6070784 nyt_200811.txt
```

行數　　詞數　　字母數

- `wc -l nyt.words`

```
1017618 nyt.words
```

為何詞數不同?（1,007,597 對 1,017,618）

Unix grep 和 wc 指令的練習

- 紐約時報 nyt_200811.txt 這個檔案，有多少大寫字？
- 有多少 4-字母的詞？
- 有多少（不同）的詞沒有母音？
  - 這些詞屬於哪一類？
- 有多少單音節的詞？
  - 只有一個母音

請區別 Type/token：types 不同詞數，tokens 詞數（不管一樣不一樣）

離線編輯命令 sed

- sed 式用來對檔案內的字串（基本上一次處理一行）做有系統的編輯（比 tr 更大更複雜的變動）
- 基本上，對每一行，但是也可以指定某一行（用正規式比對，或指定編號）
- 例如，把所有出現 George 的地方一律改為 Jane
  - `sed 's/George/Jane/' nyt_200811.txt | less`

Unix 的 sed 練習

- Count frequency of word initial consonant sequences
  - Take tokenized words
  - Delete the first vowel through the end of the word
  - Sort and count

- Count word final consonant sequences

shuf

- Randomly permutes (shuffles) the lines of a file

- **Exercises**
  - Print 10 random word tokens from the NYT excerpt
    - 10 instances of words that appear, each word instance (word token) equally likely
  - Print 10 random word types from the NYT excerpt
    - 10 different words that appear, each different word (word type) equally likely

切分欄位的 cut 指令 (針對 tab/blank 分欄檔案)

資料下載：https://www.clips.uantwerpen.be/conll2000/chunking/

相關論文：

```
head chunk.conll

Confidence NN B-NP
in IN B-PP
the DT B-NP
pound NN I-NP
is VBZ B-VP
widely RB I-VP
expected VBN I-VP
to TO I-VP
take VB I-VP
another DT B-NP
```

續

- 計算詞性標籤的頻率
- `cut -f2 -d' ' chunk.conll | sort | uniq -c | sort -nr | head`

- 計算詞、詞性
- `cut -f1,2 -d' ' chunk.conll | sort | uniq -c | head -8000 | tail`

- 如果是逗點隔開的檔案，那就用 cut –d,

Unix cut 命令的練習

- 比較 that 用為冠詞 determiner (DT) 例如 that rabbit，以及用於名詞子句 IN，如 I know that they are plastic 或關係代名詞 relative (WDT) 如 The class that I love”
  - 提示: 在 grep 時，用分隔符號（如 \t 代表 tab）

- 有哪些冠詞 determiners? 最常見的 5 是哪些?