商管程式設計（106-1）

作業六

作業設計：盧信銘
國立台灣大學資管系

截止時間：2017 年 11 月 6 日凌晨 1 點
測資助教：彭毅軒

作業繳交請至 PDOGS (http://pdogs.ntu.im/judge/)。為第一題上傳兩分 Python 3.6 原始碼 (分 case1 與 case2)。為第二、三題上傳一份 Python 3.6 原始碼 (以複製貼上原始碼的方式上傳)。作業自己做。嚴禁抄襲。不接受紙本繳交，不接受遲交。請以英文或中文作答。

除了課本的內容外，Python 線上文件也很有用: https://docs.python.org/3/。好的程式設計師會把線上文件摸熟。

除下面所述之規定外，你可以使用任何 Python 內建的 Function 與 Library。但你的實作需依照題目的說明。如果你的結果正確，但沒有依照題目規定的方式實作，則不予計分。本次作業各題之實作規定如下：第一題禁用現成的 CSV 處理模組，如 csv。第三題禁止使用 re 函式庫 (regular expression)。

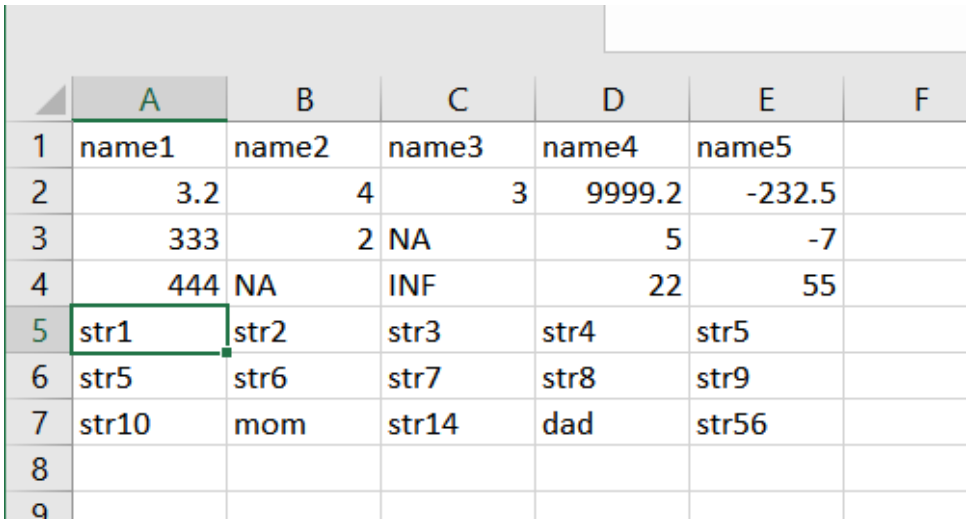
第一題

(40 points; 10 points for Case 1, 30 points for Case 2) Comma-separated values (CSV) 是一個很常使用的資料格式。這個資料格式基本上是把每一行中的資料欄位以逗點分隔，而文件本身是純文字的格式。大部分的資料處理軟體與平台，如 Excel 或 R，都支援這個格式。如果你對 CSV 格式不熟悉，可以打開 Excel，輸入一些資料之後另存成 CSV 格式，然後再用 Notepad++打開。兩相比較可以增進你對這個資料格式的了解。

本次作業將讓大家熟悉一些 CSV 檔案處理的基本議題。完整的 CSV 格式說明請參考線上文件 (https://tools.ietf.org/html/rfc4180#page-2)。然而，完全依照這個格式對初學者而言可能太複雜，因此我們將在本次作業做適度的簡化。CSV 處理主要的難點在於如何處理資料本身就包含逗點、引號、甚至換行的狀況。為簡化問題，我們主要考慮下面兩個較簡單的狀況。

Case 1. 每一個欄位的資料可以有數字或文字。資料中不能有雙引號 (") 或逗號。這是最單純的狀況，因為每一個逗號都是用來分隔資料欄位。 舉例而言，下面文字框中的 CSV 檔案內容對應到 Fig 1 中的 Excel 內容。

```
name1,name2,name3,name4,name5
3.2,4,3,9999.2,-232.5
333,2,NA,5,-7
444,NA,INF,22,55
str1,str2,str3,str4,str5
str5,str6,str7,str8,str9
str10,mom,str14,dad,str56
```



	A	B	C	D	E	F
1	name1	name2	name3	name4	name5	
2	3.2	4	3	9999.2	-232.5	
3	333	2	NA	5	-7	
4	444	NA	INF	22	55	
5	str1	str2	str3	str4	str5	
6	str5	str6	str7	str8	str9	
7	str10	mom	str14	dad	str56	
8						
9						

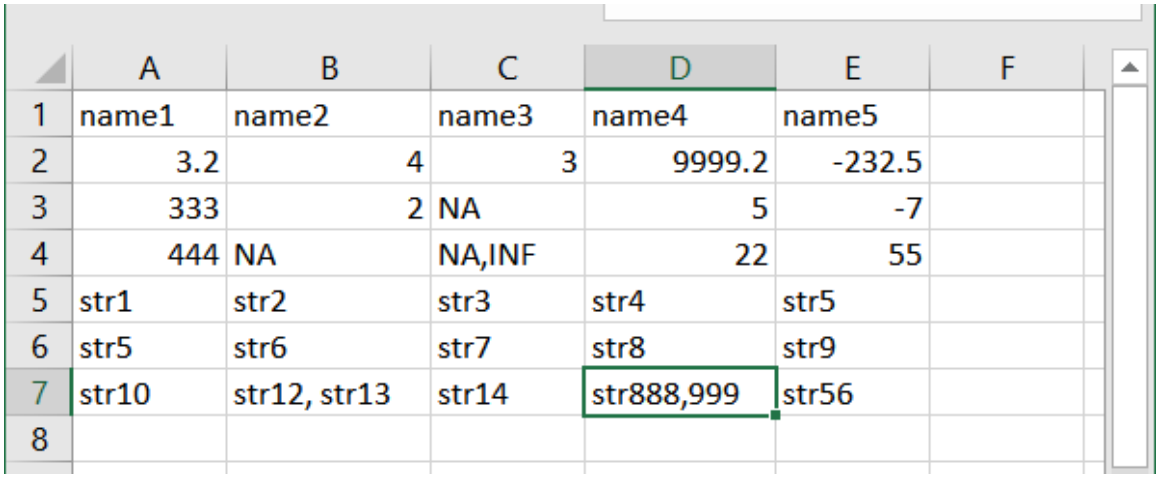
Fig 1. Example Excel Spreadsheet for Case 1.

Case 2. 每個欄位可以有數字或文字。文字中可以有逗點 (,) 但不能有雙引號 (") 或其他特殊字元 (如換行)。如果資料中有逗點，則這個欄位必須用雙引號框起來。這樣我們就可以把雙引號中的逗點解釋成資料，而雙引號之外的逗點解釋成分隔資料的符號。如果這個資料沒有逗點，則不需要以雙引號框起來。

下面的文字框中是一個例子，對應到 Fig 2 中 Excel 的資料。這個資料表中的某些位置 (C4, D7, B7) 的資料包含逗點。當你把這個資料表另存成 CSV 檔案時，這些資料就會用雙引號框起來。比如說，C4 位置的資料會被轉換成 "NA,INF" 。如此我們就知道引號中的逗點是資料的一部分，不應該被當成是分隔符號。

```
name1,name2,name3,name4,name5
3.2,4,3,9999.2,-232.5
333,2,NA,5,-7
444,NA,"NA,INF",22,55
```

```
str1,str2,str3,str4,str5
str5,str6,str7,str8,str9
str10,"str12, str13",str14,"str888,999",str56
```



	A	B	C	D	E	F
1	name1	name2	name3	name4	name5	
2	3.2	4	3	9999.2	-232.5	
3	333	2	NA	5	-7	
4	444	NA	NA,INF	22	55	
5	str1	str2	str3	str4	str5	
6	str5	str6	str7	str8	str9	
7	str10	str12, str13	str14	str888,999	str56	
8						

Fig 2. Example Excel Spreadsheet for Case 2.

(a) 請寫一個程式處理 Case 1 狀況的資料。資料處理以行 (a line) 為單位。你的程式會由使用者讀入一行的 CSV 資料，然後你的程式需要將資料切割成一個個欄位輸出，輸出時每一行是一個資料欄位。下面為一個輸出入的範例：

Sample input:
```
3.2,4,3,9999.2,232.5
```

Sample output:
```
3.2
4
3
9999.2
232.5
```

(b) 請寫一個程式處理 Case 2 狀況的資料。資料處理以行 (a line) 為單位。你的程式會由使用者讀入一行的 CSV 資料，然後你的程式需要將資料切割成一個個欄位輸出，輸出時每一行是一個資料欄位。下面為一個輸出入的範例：

Sample input:
```
str10,"str12, str13",str14,"str888,999",str56
```

Sample output:
```
str10
"str12, str13"
str14
```

```
"str888,999"
str56
```

Case 1 評估: 依據程式運算的正確性給分。一筆測試資料佔 2 分。

Case 2 評估: 本題總分中的 20 分會依據程式運算的正確性給分。一筆測試資料佔 2 分。剩下的 10 分會根據你所寫的程式品質來給分。助教會打開你的程式碼並檢閱你的程式的運算邏輯、可讀性，以及可擴充性。請寫一個「好」的程式吧！

第二題

(30 points) 本題將練習程式輸出格式的一些技巧。下面這個框框列出了環保署公布的空氣品質指標。其中第一列 (SO2,CO,…) 是標頭，後面各列則是資料。各資料欄位之間是以逗點分隔。

```
SO2,CO,O3,PM10,PM2.5,NO2,WindSpeed,WindDirec,FPMI,NOx,NO,PublishTime
2.5,0.26,13,42,8,8.1,3.1,211,1,13.32,5.25,9/9/2016 10:00
1.3,,15,18,5,4.1,1.4,98,1,6.53,2.39,9/9/2016 10:00
1.2,0.27,17,15,5,8.3,1.9,159,1,13,4.67,9/9/2016 10:00
9.2,0.44,28,82,41,15,1,340,3,19.15,4.48,9/9/2016 10:00
2.5,0.32,57,37,20,4.8,3.3,56,2,6.08,1.24,9/9/2016 10:00
1.9,0.24,16,28,13,4.9,0.9,111,2,8.2,3.35,9/9/2016 10:00
7.8,0.93,25,37,15,35,0.6,239,1,40.7,5.54,9/9/2016 10:00
8,1.29,10,30,11,25,1.2,94,1,52.06,26.94,9/9/2016 10:00
2.1,0.14,20,18,5,5.2,1.2,255,1,7.2,1.97,9/9/2016 10:00
5.2,1.73,6.7,54,25,28,1.1,187,2,69.54,41.78,9/9/2016 10:00
```

在這個情境，資料共有十二個欄位。請寫一個程式，由 input() 取的一行資料，格式如上面範例中第二列至第十一列，並產製”好看的輸出”。好看的輸出定義如下：

1. 第一欄至第十一欄每欄的寬度為五個字元，第十二欄的寬度為 16 個字元。
2. 欄跟欄中間隔一個空白 (半形)。
3. 欄位內容如果超過指定長度 n，則取前面 n 個字元。
4. 缺值應輸出 “-“，且應向中對齊。偶數寬度則讓前半部短一個字元。
5. SO2, NO2, WindSpeed：輸出浮點數，取一位小數，四捨五入。整數部分不足三位者前面補 0。小數部分為零者須輸出 0。
6. CO, NOx, NO：輸出浮點數，取兩位小數，四捨五入。整數部分不足兩位者前面補 0。小數部分為零者須輸出 0。
7. O3, PM10, PM2.5, WindDirec, FPMI：輸出整數，最多五位數，向右對齊。前面不補零。
8. PublishTime：照原始字串輸出，向右對齊。

範例輸入：

```
2.5,0.26,13,42,8,8.1,3.1,211,1,13.32,5.25,9/9/2016 10:00
```

範例輸出：

```
002.5 00.26    13    42     8 008.1 003.1   211     1 13.32 05.25   9/9/2016 10:00
```

範例輸入：

```
2.1,0.14,20,18,5,5.2,1.2,255,1,7.2,1.97,9/9/2016 10:00
```

範例輸出：

```
002.1 00.14    20    18     5 005.2 001.2   255     1 07.20 01.97   9/9/2016 10:00
```

本題分數都由程式運算的正確性給分，一筆測試資料佔 2 分。

第三題

(30 points) 我們在課程中已經熟悉了如何用 split()將一個字串切割成多個小的字串。然而 split()只能選定一個切分字串，比如說：

```
a="hello, my friend"
a.split(' ')
```

上面的例子是用空白切割。所以結果會是 ['hello,', 'my', 'friend']。在很多的應用中，我們需要將一個輸入的文本切割成一個一個的單字。在上面的例子，合理的輸出應該是 ['hello', 'my', 'friend']，因為逗點”,”也應該算是分隔字元。如果輸入的字串是“economic, flat, high, moderate, (weakening, weakened), mixed”，那輸出應該是所有出現過的單字: [‘economic’, ‘flat’, ‘high’, ‘moderate’, ‘weakening’, ‘weakened’, ‘mixed’]。

寫一個 Python 程式，利用 input() 讀入一段文字，並輸入順序輸出所有出現過的單字，每個單字輸出一行，所有的單字應轉成小寫。單字的定義是所有被分隔字元隔開的字串。本題所使用的分隔字元是：

```
sepchar = ' ,\'";.!:()[]{}\n\r\t=+/\\><'
```

任何在 spechar 中的字元都是分隔字元。

注意不要輸出額外的空白或其他字元。

範例輸入:

```
Growth of consumer spending ranged from slight to moderate in most Districts, while auto sales were somewhat mixed, as activity has begun to drop off from previously high levels in some Districts.
```

範例輸出:

```
growth
of
consumer
spending
ranged
from
slight
to
moderate
in
most
districts
while
auto
sales
were
somewhat
mixed
as
activity
has
begun
to
drop
off
from
previously
high
levels
in
some
districts
```

範例輸入:

```
ind2 = abs(allt2)> 1.5 allt3_agent = allt2[ind2] nfeat2_agent = length(allt3_agent) cat("Number of features from user_agent",  "=", nfeat2_agent, "\n") nextcol = ncol(rtb2)+1
```

範例輸出:

```
ind2
abs
allt2
1
5
allt3_agent
allt2
ind2
nfeat2_agent
length
allt3_agent
cat
number
of
features
from
user_agent
nfeat2_agent
n
nextcol
ncol
rtb2
1
```

範例輸入:

```
Cindy & I are keeping America's sailors aboard the USS John S McCain in our prayers tonight - appreciate the work of search & rescue crews https://t.co/jzk9giXbfg
```

範例輸出:

```
cindy
&
i
are
keeping
america
s
sailors
aboard
the
uss
```

john
s
mccain
in
our
prayers
tonight
-
appreciate
the
work
of
search
&
rescue
crews
https
t
co
jzk9gixbfg

本題分數都由程式運算的正確性給分，一筆測試資料佔 2 分。