Embarak _Ch04_File IO Processing _ Regular Expressions

September 5, 2018

1 Ch04 File processing and Regular expressions

2 File processing

```
In [2]: Name = input ("Enter your name: ")
        Name

Enter your name: Osama Hashim


Out[2]: 'Osama Hashim'

In [3]: Mark = input("Enter your mark: ")
        Mark = float(Mark)

Enter your mark: 92


In [4]: print ("Welcome to Grading System \nHCT 2018")
        print ("\nCampus\t Name\t\tMark\tGrade")
        if (Mark>=85):
            Grade="B+"
        print ("FMC\t", Name,"\t",Mark,"\t", Grade)

Welcome to Grading System
HCT 2018

Campus          Name                 Mark         Grade
FMC         Osama Hashim           92.0           B+
```

2.0.1 Files attributes

```
In [41]: # Open a file and find its attributes
         Filehndl = open("Egypt.txt", "r")
         print ("Name of the file: ", Filehndl.name)
         print ("Closed or not : ", Filehndl.closed)
         print ("Opening mode : ", Filehndl.mode)
```

```
Name of the file:  Egypt.txt
Closed or not :  False
Opening mode :  r
```

2.0.2 Open and close files

```
In [40]: Filehndl = open("Egypt.txt", "r")
         print ("Closed or not : ", Filehndl.closed)
         Filehndl.close()
         print ("Closed or not : ", Filehndl.closed)
```

```
Closed or not :  False
Closed or not :  True
```

```
In [39]: Filehndl = open("Egypt.txt", "w+")
         Filehndl.write( "Python Processing Files\nMay 2018!!\n")

         # Close opend file
         Filehndl.close()
```

2.0.3 Rename and delete files

```
In [34]: import os
         os.rename( "Egypt.txt", "test2.txt" )
         os.remove( "test2.txt" )
```

2.1 Directories in Python

```
In [ ]: import os
        os.mkdir("Data 1")   # create a directory
        os.mkdir("Data_2")
        os.chdir("Data_2")    # create a childe directory
        os.getcwd()          # Get the current working directory

        os.rmdir('Data 1')   # remove a directory
        os.rmdir('Data_2')   # remove a directory
```

```
In [44]: import os
         os.getcwd()   # Get the current working directory
```

```
Out[44]: '/home/nbuser/library'
```

```
In [43]: os.chdir('/home/nbuser/library')
```

2.2 open and process files

```
In [45]: print("\nSearching Through a File\n")
         fhand = open('Emails.txt')
         for line in fhand:
             line = line.rstrip()
             if line.startswith('From:') :
                 print (line)
```

```
Searching Through a File

From: stephen.marquard@uct.ac.za
From: louis@media.berkeley.edu
From: zqian@umich.edu
From: rjlowe@iupui.edu
From: zqian@umich.edu
From: rjlowe@iupui.edu
From: cwen@iupui.edu
From: cwen@iupui.edu
From: gsilver@umich.edu
From: gsilver@umich.edu
From: zqian@umich.edu
From: gsilver@umich.edu
From: wagnermr@iupui.edu
From: zqian@umich.edu
From: antranig@caret.cam.ac.uk
From: gopal.ramasammycook@gmail.com
From: david.horwitz@uct.ac.za
From: david.horwitz@uct.ac.za
From: david.horwitz@uct.ac.za
From: david.horwitz@uct.ac.za
From: stephen.marquard@uct.ac.za
From: louis@media.berkeley.edu
From: louis@media.berkeley.edu
From: ray@media.berkeley.edu
From: cwen@iupui.edu
From: cwen@iupui.edu
From: cwen@iupui.edu
```

```
In [46]: print ("\nUsing in to select lines  // only print lines which has specific string ")
         fhand = open('Emails.txt')
         for line in fhand:
             line = line.rstrip()
             if not '@uct.ac.za' in line :
                 continue
             print (line)
```

```
Using in to select lines  // only print lines which has specific string
From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008
X-Authentication-Warning: nakamura.uits.iupui.edu: apache set sender to stephen.marquard@uct.ac.
From: stephen.marquard@uct.ac.za
Author: stephen.marquard@uct.ac.za
From david.horwitz@uct.ac.za Fri Jan  4 07:02:32 2008
X-Authentication-Warning: nakamura.uits.iupui.edu: apache set sender to david.horwitz@uct.ac.za
From: david.horwitz@uct.ac.za
Author: david.horwitz@uct.ac.za
r39753 | david.horwitz@uct.ac.za | 2008-01-04 13:05:51 +0200 (Fri, 04 Jan 2008) | 1 line
From david.horwitz@uct.ac.za Fri Jan  4 06:08:27 2008
X-Authentication-Warning: nakamura.uits.iupui.edu: apache set sender to david.horwitz@uct.ac.za
From: david.horwitz@uct.ac.za
Author: david.horwitz@uct.ac.za
From david.horwitz@uct.ac.za Fri Jan  4 04:49:08 2008
X-Authentication-Warning: nakamura.uits.iupui.edu: apache set sender to david.horwitz@uct.ac.za
From: david.horwitz@uct.ac.za
Author: david.horwitz@uct.ac.za
From david.horwitz@uct.ac.za Fri Jan  4 04:33:44 2008
X-Authentication-Warning: nakamura.uits.iupui.edu: apache set sender to david.horwitz@uct.ac.za
From: david.horwitz@uct.ac.za
Author: david.horwitz@uct.ac.za
From stephen.marquard@uct.ac.za Fri Jan  4 04:07:34 2008
X-Authentication-Warning: nakamura.uits.iupui.edu: apache set sender to stephen.marquard@uct.ac.
From: stephen.marquard@uct.ac.za
Author: stephen.marquard@uct.ac.za
```

```
In [47]: print("\nSearching Through a File\n")
         fhand = open('Emails.txt')
         for line in fhand:
             line = line.rstrip()
             if line.startswith('From:') :
                 line = line.split()
                 print (line[1])
```

```
Searching Through a File

stephen.marquard@uct.ac.za
louis@media.berkeley.edu
zqian@umich.edu
rjlowe@iupui.edu
zqian@umich.edu
rjlowe@iupui.edu
cwen@iupui.edu
cwen@iupui.edu
```

```
gsilver@umich.edu
gsilver@umich.edu
zqian@umich.edu
gsilver@umich.edu
wagnermr@iupui.edu
zqian@umich.edu
antranig@caret.cam.ac.uk
gopal.ramasammycook@gmail.com
david.horwitz@uct.ac.za
david.horwitz@uct.ac.za
david.horwitz@uct.ac.za
david.horwitz@uct.ac.za
stephen.marquard@uct.ac.za
louis@media.berkeley.edu
louis@media.berkeley.edu
ray@media.berkeley.edu
cwen@iupui.edu
cwen@iupui.edu
cwen@iupui.edu
```

2.3 Regular Expressions

```
In [48]: import re
         print ("\nRegular Expressions\n'^X.*:' \n")
         hand = open('Data.txt')
         for line in hand:
             line = line.rstrip()
             y = re.findall('^X.*:',line)
             print (y)
```

```
Regular Expressions
'^X.*:'

['X-Sieve:']
['X-DSPAM-Result:']
['X-DSPAM-Confidence:']
['X- Content-Type-Message-Body:']
['X-Plane is behind schedule:']
```

```
In [49]: print ("\nRegular Expressions\nWild-Card Characters  '^X-\S+:'\n")
         hand = open('Data.txt')
         for line in hand:
             line = line.rstrip()
             y = re.findall('^X-\S+:',line) # match any non white space characters
             print (y)
```

```
Regular Expressions
Wild-Card Characters  '^X-\S+:'

['X-Sieve:']
['X-DSPAM-Result:']
['X-DSPAM-Confidence:']
[]
[]
```

```
In [50]: print ("\n Matching and Extracting Data \n")
         x = 'My 2 favorite numbers are 19 and 42'
         y = re.findall('[0-9]+',x)
         print (y)
```

```
 Matching and Extracting Data

['2', '19', '42']
```

```
In [51]: y = re.findall('[AEsOUn]+',x)  # find any of these characters in string
         print (y)

['n', 's', 'n']
```

```
In [52]: print ("\nGreedy Matching \n")
         x = 'From: Using the : character'
         y = re.findall('^F.+:', x)
         print (y)
```

```
Greedy Matching

['From: Using the :']
```

```
In [53]: print ("\nNon-Greedy Matching \n")
         x = 'From: Using the : character'
         y = re.findall('^F.+?:', x)
         print (y)
```

```
Non-Greedy Matching

['From:']
```

```
In [54]: import re
         print ("\nFine-Tuning String Extraction \n")
         mystr="From ossama.embarak@hct.ac.ae Sat Jun 5 08:14:16 2018"
         Extract  = re.findall('\S+@\S+',mystr)
         print (Extract)
         E_xtracted = re.findall('^From.*? (\S+@\S+)',mystr) # non greedy white space
         print (E_xtracted)
         print (E_xtracted[0])
```

```
Fine-Tuning String Extraction

['ossama.embarak@hct.ac.ae']
['ossama.embarak@hct.ac.ae']
ossama.embarak@hct.ac.ae
```

```
In [57]: mystr="From ossama.embarak@hct.ac.ae Sat Jun 5 08:14:16 2018"
         atpos = mystr.find('@')
         sppos = mystr.find(' ',atpos) # find white space starting from atpos
         host = mystr[atpos+1 : sppos]
         print (host)
         usernamepos =mystr.find(' ')
         username = mystr[usernamepos+1 : atpos]
         print (username)
```

```
hct.ac.ae
ossama.embarak
```

```
In [58]: print ("\n The Regex Version\n")
         import re
         mystr="From ossama.embarak@hct.ac.ae Sat Jun 5 08:14:16 2018"
         Extract = re.findall('@([^ ]*)',mystr)
         print (Extract)
         Extract = re.findall('^From .*@([^ ]*)',mystr)
         print (Extract)
```

```
 The Regex Version

['hct.ac.ae']
['hct.ac.ae']
```

```
In [59]: print ("\nScape character \n")
         mystr = 'We just received $10.00 for cookies and $20.23 for juice'
         Extract = re.findall('\$[0-9.]+',mystr)
         print (Extract)
```

```
Scape character

['$10.00', '$20.23']
```

2.4 Exercises

```
In [60]: import re
         CoursesData = """101   COM   Computers
         205   MAT   Mathematics
         189   ENG    English"""

In [61]: # Extract all course numbers
         Course_numbers = re.findall('[0-9]+', CoursesData)
         print (Course_numbers)

         # Extract all course codes
         Course_codes = re.findall('[A-Z]{3}', CoursesData)
         print (Course_codes)

         # Extract all course names
         Course_names = re.findall('[A-Za-z]{4,}', CoursesData)
         print (Course_names)

['101', '205', '189']
['COM', 'MAT', 'ENG']
['Computers', 'Mathematics', 'English']


In [62]: # compile the regex and search the pattern
         regex_num = re.compile('\d+')
         s = regex_num.search(CoursesData)

         print('Starting Position: ', s.start())
         print('Ending Position: ', s.end())
         print(CoursesData[s.start():s.end()])

Starting Position:  0
Ending Position:  3
101


In [63]: # define the course text pattern groups and extract
         course_pattern = '([0-9]+)\s*([A-Z]{3})\s*([A-Za-z]{4,})'
         re.findall(course_pattern, CoursesData)

Out[63]: [('101', 'COM', 'Computers'),
          ('205', 'MAT', 'Mathematics'),
          ('189', 'ENG', 'English')]
```

```
In [64]: print(re.findall('[a-zA-Z]+', CoursesData))  # [] Matches any character inside
```

```
['COM', 'Computers', 'MAT', 'Mathematics', 'ENG', 'English']
```

```
In [65]: print(re.findall('[0-9]+', CoursesData))  # [] Matches any character inside
```

```
['101', '205', '189']
```

```
In [66]: import re
         CoursesData = """10   COM   Computers
         205   MAT   Mathematics
         1899   ENG    English"""
         print(re.findall('\d{4}', CoursesData))  # {n} Matches repeat n times.
         print(re.findall('\d{2,4}', CoursesData))
```

```
['1899']
['10', '205', '1899']
```