LIVE 4: Strings and Regex

- Focus: Basics of strings and regex in Python + Simple problem solving.
- Prereq: Basic knowledge of Strings and Regex in Python + previous code-sessions.
- Reference for basics:
  - https://docs.python.org/3/howto/regex.html (https://docs.python.org/3/howto/regex.html)
  - https://docs.python.org/3/library/re.html (https://docs.python.org/3/library/re.html)
  - https://www.w3schools.com/python/python_strings.asp (https://www.w3schools.com/python/python_strings.asp)
  - https://www.geeksforgeeks.org/python-strings/ (https://www.geeksforgeeks.org/python-strings/)

Quick recap of Regex in Python

- Go through multiple examples to understand regex better
- Key life-skill: learn from resources on the internet.
- https://docs.python.org/3/howto/regex.html (https://docs.python.org/3/howto/regex.html)
- https://www.w3schools.com/python/python_regex.asp (https://www.w3schools.com/python/python_regex.asp)
- https://www.tutorialspoint.com/python/python_reg_expressions.htm (https://www.tutorialspoint.com/python/python_reg_expressions.htm)

Problem-1: Mask personal information in email and phone numbers

- Email: xxxxxxxxx@aaaa.zzzz
  - Masked:x#####x@aaaa.zzzz [FIVE # between first and last char of the name]

- Phone: digits 0-9 or any of the characters from { '-', '(', ')', ' '}
  - Example: 1(234)567-890 --> ###-###-7890"

In [13]:

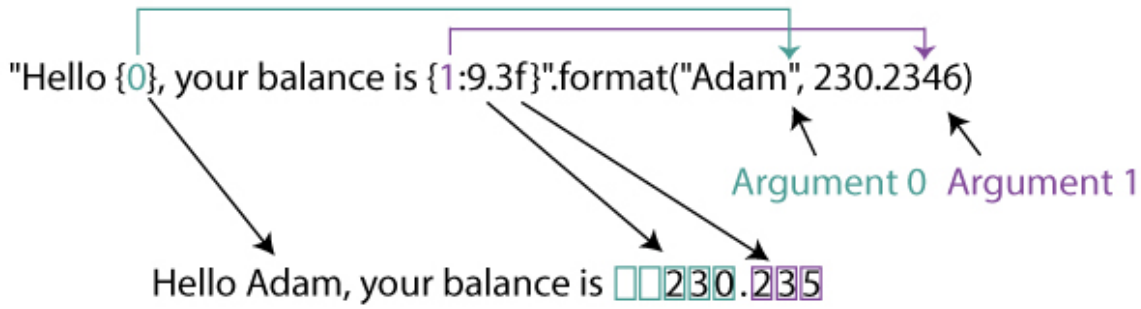
```
# email-masking
s = "abcd@efgh.com";

#output: a#####d@efgh.com
# Any suggestions?
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#

#simple string formatting : https://www.programiz.com/python-programming/methods/string/format

from IPython.display import Image
Image(url= "https://cdn.programiz.com/sites/tutorial2program/files/python-format-positional-argument.jpg")
```

Out[13]:



In [9]:

```
# boundary case: check if email or not
s = "abcd@efgh.com";

def maskEmail(s):
    if '@' in s:
        name, domain = s.split('@')
        return ("{0}#####{1}@{2}".format(name[0], name[-1], domain));

print(maskEmail(s))
```

```
a#####d@efgh.com
```

In [10]:

```
# BOUDNARY CASE: a@bcdef.com
print(maskEmail("a@bcdef.com"))
```

```
a#####a@bcdef.com
```

In [11]:

```
# BOUDNARY CASE: abcd.com
print(maskEmail("abcd.com"))
```

```
None
```

In [12]:

```
# BOUDNARY CASE: abcd@cdef
print(maskEmail("abcd@cdef"))
```

```
a#####d@cdef
```

In [117]:

```
# Check if email is valid is another function.
# Any suggestions?
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
import re
def isValidEmail(s):
    #refer:https://www.w3schools.com/python/python_regex.asp for regex syntax
    #https://docs.python.org/2/library/re.html

    res = re.search('^\w+([\.-]?\w+)*@\w+([\.-]?\w+)*(\.\w{2,3})+$', s, re.IGNOR
ECASE)    #https://www.geeksforgeeks.org/check-if-email-address-valid-or-not-in-p
ython/
    print(res)

    if(res):
        return True;
    else:
        return False;

print(isValidEmail("abcd@cdef"))
```

```
None
False
```

In [118]:

```
print(isValidEmail("abcd@cdef.c"))
```

```
None
False
```

In [65]:

```
print(isValidEmail("a@cdef.com"))
```

```
<re.Match object; span=(0, 10), match='a@cdef.com'>
True
```

In [66]:

```
print(isValidEmail("abcd@iisc.ac.in"))
```

```
<re.Match object; span=(0, 15), match='abcd@iisc.ac.in'>
True
```

In [67]:

```
regex = '^\w+([\.-]?\w+)*@\w+([\.-]?\w+)*(\.\w{2,3})+$' # highly non-readbale co
de

#https://docs.python.org/2/library/re.html
regex_verbose = re.compile(r"""                    # VERY readable and easy to und
erstand. Software maintanability.
            ^\w+([\.-]?\w+)*              # start, \w+,
            @                             # single @ sign
            \w+([\.-]?\w+)*               # Domain name
            (\.\w{2,3})+$                 # .com, .ac.in,
             """,re.VERBOSE | re.IGNORECASE)


res = regex_verbose.match("abcd@iisc.ac.in");

print(res.string)
print(res)
```

```
abcd@iisc.ac.in
<re.Match object; span=(0, 15), match='abcd@iisc.ac.in'>
```

In [78]:

```
# PHONE NUMBER MASKING
#Example: 1(234)567-890 --> ###-###-7890"


ph = "1(234)567-890"


digits = re.sub("\D", "", ph)  # \D=>non-decimal, re.substitute, https://docs.python.org/3/library/re.html

print(digits)

masked =  "###-###-{}".format(digits[-4:])
print(masked)
```

```
1234567890
###-###-7890
```

In [82]:

```
def maskPhoneNum(ph):
    digits = re.sub("\D", "", ph)  # \D=>non-decimal, re.substitute, https://docs.python.org/3/library/re.html
    if len(digits) != 10: # BOUNDARY CASE
        return None;
    else:
        masked =  "###-###-{}".format(digits[-4:])
        return masked

print(maskPhoneNum("1(234)567-890"))
```

```
###-###-7890
```

print(maskPhoneNum("1(234)567-89"))

Exercise: 12 digit phone numbers with 2 digits of ISD code strtaing with +

- e.g: +86-(99)12345678 ----> (+86)-###-###-5678

Problem 2: Extract data from a PDF invoice

- Given a PDF [https://slicedinvoices.com/pdf/wordpress-pdf-invoice-plugin-sample.pdf (https://slicedinvoices.com/pdf/wordpress-pdf-invoice-plugin-sample.pdf)], extract predefined key fields from this PDF
- Assume the format is fixed.

NOTE: Download and save the above PDF as invoice.pdf in the same folder as your iPython notebook for the following code to work

In [85]:

```
# https://realpython.com/pdf-python/#history-of-pypdf-pypdf2-and-pypdf4

!pip3 install pyPDF4
```

```
Collecting pyPDF4
  Downloading https://files.pythonhosted.org/packages/4f/1f/509b4485
0c475c101aa5b5c9b81755cedd363389d6fbb5c53be6fa915a61/PyPDF4-1.27.0.t
ar.gz (63kB)
    100% |████████████████████████████████| 71kB 1.5MB/s ta 0:00:011
Building wheels for collected packages: pyPDF4
  Building wheel for pyPDF4 (setup.py) ... done
  Stored in directory: /Users/varma/Library/Caches/pip/wheels/eb/4f/
15/c64d533cb496fd874f56045fe30e8cc0ac59f99ecdd718040d
Successfully built pyPDF4
Installing collected packages: pyPDF4
Successfully installed pyPDF4-1.27.0
```

In [87]:

```
# Google "pyPDF extract text" ---> https://www.soudegesu.com/en/post/python/extract-text-from-pdf-with-pypdf2/
import PyPDF4

FILE_PATH = './invoice.pdf'

with open(FILE_PATH, mode='rb') as f:
    reader = PyPDF4.PdfFileReader(f)
    page = reader.getPage(0)
    print(page.extractText())
```

```
Invoice
Payment is due within 30 days from date of invoice. Late payment is
subject to fees of 5% per month.
Thanks for choosing
DEMO - Sliced Invoices
 |
admin@slicedinvoices.com
Page 1/1
From:
DEMO - Sliced Invoices
Suite 5A-1204
123 Somewhere Street
Your City AZ 12345
admin@slicedinvoices.com
Invoice Number
INV-3337
Order Number
12345
Invoice Date
January 25, 2016
Due Date
January 31, 2016
Total Due
$93.50
To:
Test Business
123 Somewhere St
Melbourne, VIC 3000
test@test.com
Hrs/Qty
Service
Rate/Price
Adjust
Sub Total
1.00
Web Design
This is a sample description...
$85.00
0.00%
$85.00
Sub Total
$85.00
Tax
$8.50
Total
$93.50
ANZ Bank
ACC # 1234 1234
BSB # 4321 432
Paid
```

In [88]:

```
import PyPDF4

FILE_PATH = './invoice.pdf'

with open(FILE_PATH, mode='rb') as f:
    reader = PyPDF4.PdfFileReader(f)
    page = reader.getPage(0)
    txt = page.extractText();
```

In [98]:

```
# extract invoice number

m = re.findall("INV-[0-9]*", txt)
print(m)
```

```
['INV-3337']
```

In [100]:

```
# extract amounts
m = re.findall("$[0-9]*\.[0-9]*", txt)
print(m)
```

```
[]
```

In [ ]:

```

```

In [104]:

```
# extract amounts
m = re.findall("\$[0-9]*\.[0-9]*", txt)
print(m)
```

```
['$93.50', '$85.00', '$85.00', '$85.00', '$8.50', '$93.50']
```

In [106]:

```
# Extract Total Due:
m = re.findall("Total Due\$[0-9]*\.[0-9]*", txt)
print(m)
```

```
[]
```

In [107]:

```
# Any suggestions?
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
# Extract Total Due:
m = re.findall("Total Due\n\$[0-9]*\.[0-9]*", txt)
print(m)
```

```
['Total Due\n$93.50']
```

In [109]:

```
print(re.findall("\$[0-9]*\.[0-9]*",m[0]))
```

```
['$93.50']
```

Ques: How do we handle cases where we want to extract data from multiple invoice formats?

Assignment: Extract email-addresses from the PDF

We will continue from here tomorrow. Pelase go through regex- referecnes in detail for tomorrow's session. We will solve a few product based company interview questions.

In [ ]: