In [ ]:

- .jpg

- .png

- where is your file located: <location>

- what is your file name

- what is the type of file (extention)

C:\Users\omkar\OneDrive\Documents\Data science\Naresh IT\Datafiles
mbox-short

.txt

```
file_location="C:\Users\omkar\OneDrive\Documents\Data science\Naresh IT\Data
file_name="mbox-short"
extention=".txt"
```

In [ ]: In [ ]:

```
file_location+file_name+extention
```

```
"C:\Users\omkar\OneDrive\Documents\Data science\Naresh IT\Datafiles"
```

In [ ]:

```
Cell In[2], line 1
 "C:\Users\omkar\OneDrive\Documents\Data science\Naresh IT\Datafiles"  ^
SyntaxError: (unicode error) 'unicodeescape' codec can't decode bytes in p osition 2-3: truncated \UXXXXXXXX escape
```

In [ ]: In [2]:

whenever you want to read any file

type of file

- .txt
- .csv (comma seperted value)
- .xlsx
- .json (dictionay format)
- .xml (IOT data)
- .parquet (encoded files)
- .delta (encoded files)
- .pdf

In [ ]:

In [ ]: In [ ]: In [3]:

- unicode error : single slash
- in order to read the file , we need provide '\\'

```
"C:\\Users\\omkar\\OneDrive\\Documents\\Data science\\Naresh IT\\Datafiles\\
```

```
"C:\\Users\\omkar\\OneDrive\\Documents\\Da
```

```
ta science\\Naresh IT\\Datafiles\\        uments\\Data science\\Naresh IT\\D
                                         file_path

file_path="C:\\Users\\omkar\\OneDrive\\Doc
```

Out[3]:
```
'C:\\Users\\omkar\\OneDrive\\Documents\\Data science\\Naresh IT\\Datafiles
\\mbox-short.txt'
```

```
                    r/write:w
                    # emcoding:
In [4]:             cp1252/utf-8
open(file_path,encod
ing='utf-8')

# mode: read:
```

Out[4]:
```
<_io.TextIOWrapper name='C:\\Users\\omkar\\OneDrive\\Documents\\Data scien
ce\\Naresh IT\\Datafiles\\mbox-short.txt' mode='r' encoding='cp1252'>
```

whenever eny file has spl charcters it will not read
that time you need to provide encoding value
different emcoded files has different encoding method
'utf-8'/'cp-1252'

```
                                         #\t: tab
In [14]:                                 y ( y ) ;
file_path="C:\\Users\\omkar\\OneDrive\\Doc  Sat, 05 Jan 2008 09:14:16 -0500
uments\\Data science\\Naresh IT\\D
file=open(file_path,encoding='utf-8-sig')
print(file.read())

#\n: new line
          Received: from holes.mr.itd.umich.edu (holes.mr.itd.umich.edu [141.211.1
          4.79])
                  by flawless.mail.umich.edu () with ESMTP id m05EEFR1013674;
                  Sat, 5 Jan 2008 09:14:15 -0500
          Received: FROM paploo.uhi.ac.uk (app1.prod.collab.uhi.ac.uk [194.35.219.
          184])
                   BY holes.mr.itd.umich.edu ID 477F90B0.2DB2F.12494 ;
                   5 Jan 2008 09:14:10 -0500
          Received: from paploo.uhi.ac.uk (localhost [127.0.0.1])
                  by paploo.uhi.ac.uk (Postfix) with ESMTP id 5F919BC2F2;
                  Sat, 5 Jan 2008 14:10:05 +0000 (GMT)
          Message-ID: <200801051412.m05ECIaH010327@nakamura.uits.iupui.edu>
          Mime-Version: 1.0
          Content-Transfer-Encoding: 7bit
          Received: from prod.collab.uhi.ac.uk ([194.35.219.182])
            by paploo.uhi.ac.uk (JAMES SMTP Server 2.1.3) with SMTP ID 899  for
           <source@collab.sakaiproject.org>;
            Sat, 5 Jan 2008 14:09:50 +0000 (GMT)

           Received: from nakamura uits iupui edu (nakamura uits iupui edu [134 68
In [ ]:
```

```
                                    full location

                                    # python file and data file both
                                    are in same location # you no need
                                    to provide the location
```

In [1]:

```
# in the above case
# your data file and python file
both are in different location #
```

In [2]:

```
import os

os.getcwd()
# get current working directory
```

Out[2]: 'C:\\Users\\omkar\\Documents'

In [ ]:

In [3]:

```
# i will copy data file in my python
location


# data file and python file both are in
same location
# file_name+<extention>
# no need to provide location

file_path="mbox-short.txt"
file=open(file_path,encoding='utf-8-sig')
print(file.read())


# **Note: data file and python file in
same location
```

```
for <source@collab.sakaiproject.org>;
Sat, 5 Jan 2008 09:12:19 - 0500
Received: (from apache@localhost)
by nakamura.uits.iupui.edu
(8.12.11.20060308/8.12.11/Submit) id
m05ECIaH010327
for source@collab.sakaiproject.org; Sat,
5 Jan 2008 09:12:18 -05 00
Date: Sat, 5 Jan 2008 09:12:18 -0500
X-Authentication-Warning:
nakamura.uits.iupui.edu: apache set
sender to stephen.marquard@uct.ac.za
using -f
To: source@collab.sakaiproject.org
From: stephen.marquard@uct.ac.za
```

```
220.122])
```

```
by shmi.uhi.ac.uk (Postfix) with ESMTP id
A215243002
for <source@collab.sakaiproject.org>;
Sat, 5 Jan 2008 14:13:33 +0000 (GMT)
Received: from nakamura.uits.iupui.edu
(localhost [127.0.0.1]) by
nakamura.uits.iupui.edu
(8.12.11.20060308/8.12.11) with ESMTP id
m05ECJVp010329
```

content/branches/sakai_2-5-x/conte

Subject: [sakai] svn commit: r39772 -

In [4]: In [6]:

```
# change directory dir(os)
'__builtins__',             os.getcwd()
'__doc__',
'__file__',
'__loader__',
'__name__',
'__package__',
'__spec__',
'_check_methods',
'_execvpe',
'_exists',
'_exit',
'_fspath',
'_get_exports_list',
'_walk',
'_wrap_close',
'abc',
'abort',
'access',
'add_dll_directory',
'altsep',
```

Out[6]: 'C:\\Users\\omkar\\Documents'

In [7]:
```
path="C:\\Users\\omkar\\OneDrive\\Documents\\Data science\\Naresh IT"
os.chdir(path)
```

In [7]:
```
os.getcwd() #
```

In [8]:
```
path="C:\\Users\\omkar\\OneDrive\\Documents\\Data science\\Naresh IT"
os.chdir(path)
```

Out[8]: 'C:\\Users\\omkar\\OneDrive\\Documents\\Data science\\Naresh IT'

In [10]:
```
# - python file

# - data file

# - both are in different location

# - you need to provide full path

# - in code you alredy changed the directory, where your data file existed


file_path="mbox-short1.txt"
file=open(file_path,encoding='utf-8-sig')
print(file.read())


#file_path="C:\\Users\\omkar\\OneDrive\\Documents\\Data science\\Naresh IT\
#file=open(file_path,encoding='utf-8-sig')
#print(file.read())
```

```
Return-Path:
<postmaster@collab.sakaiproject.org>
Received: from murder (mail.umich.edu
[141.211.14.90])
 by frankenstein.mail.umich.edu (Cyrus
v2.3.8) with LMTPA;  Sat, 05 Jan 2008
09:14:16 -0500
X-Sieve: CMU Sieve 2.3
Received: from murder ([unix socket])
 by mail.umich.edu (Cyrus v2.2.12) with
LMTPA;
 Sat, 05 Jan 2008 09:14:16 -0500
Received: from holes.mr.itd.umich.edu
(holes.mr.itd.umich.edu [141.211.1 4.79])
by flawless.mail.umich.edu () with ESMTP
id m05EEFR1013674; Sat, 5 Jan 2008
09:14:15 -0500
Received: FROM paploo.uhi.ac.uk
(app1.prod.collab.uhi.ac.uk [194.35.219.
184])
BY holes.mr.itd.umich.edu ID
477F90B0.2DB2F.12494 ;
 5 Jan 2008 09:14:10 -0500
Received: from paploo.uhi.ac.uk (localhost
[127.0.0.1])
by paploo.uhi.ac.uk (Postfix) with ESMTP
id 5F919BC2F2;
```

In [11]:

In [12]:

```
From stephen.marquard@uct.ac.za Sat Jan 5
09:14:16 2008
```

```
S t 5 J 2008 14 10 05 +0000 (GMT)
```

```
path="C:\\Users\\omkar\\Documents"
os.chdir(path)
```

```
os.getcwd()
```

```
Out[12]: 'C:\\Users\\omkar\\Documents'
In [13]: In [14]:
```

```
FileNotFoundError Traceback (most recent
call las t)
Cell In[13], line 2
 1 file_path="mbox-short1.txt"
----> 2
file=open(file_path,encoding='utf-8-sig')
 3 print(file.read())

File
~\anaconda3\Lib\site-packages\IPython\core
\interactiveshell.py:286, i n
_modified_open(file, *args, **kwargs)
 279 if file in {0, 1, 2}:
 280 raise ValueError(
 281 f"IPython won't let you open
fd={file} by default "  282 "as it is
likely to crash IPython. If you know what
you ar e doing, "
 283 "you can use builtins' open."
 284 )
--> 286 return io_open(file, *args,
**kwargs)

FileNotFoundError: [Errno 2] No such file
or directory: 'mbox-short1.txt'
```

```
file_path="mbox-short1.txt"
file=open(file_path,encoding='utf-8-sig')
print(file.read())
```

```
file_path="C:\\Users\\omkar\\OneDrive\\Doc
uments\\Data science\\Naresh IT\\m
file=open(file_path,encoding='utf-8-sig')
print(file.read())
```

```
---------------------------------------
------------------------------ -
```

```
From stephen.marquard@uct.ac.za Sat Jan 5 09:14:16 2008

Return-Path: <postmaster@collab.sakaiproject.org>
Received: from murder (mail.umich.edu [141.211.14.90])
         by frankenstein.mail.umich.edu (Cyrus v2.3.8) with LMTPA;
        Sat, 05 Jan 2008 09:14:16 -0500
X-Sieve: CMU Sieve 2.3
Received: from murder ([unix socket])
         by mail.umich.edu (Cyrus v2.2.12) with LMTPA;
         Sat, 05 Jan 2008 09:14:16 -0500
Received: from holes.mr.itd.umich.edu (holes.mr.itd.umich.edu [141.211.1
4.79])
        by flawless.mail.umich.edu () with ESMTP id m05EEFR1013674;
        Sat, 5 Jan 2008 09:14:15 -0500
Received: FROM paploo.uhi.ac.uk (app1.prod.collab.uhi.ac.uk [194.35.219.
184])
         BY holes.mr.itd.umich.edu ID 477F90B0.2DB2F.12494 ;
         5 Jan 2008 09:14:10 -0500
Received: from paploo.uhi.ac.uk (localhost [127.0.0.1])
        by paploo.uhi.ac.uk (Postfix) with ESMTP id 5F919BC2F2;
```

Sat 5 Jan 2008 14:10:05 +0000 (GMT)

In [ ]:

- now i will not change my data file

- os.chdir(<path>)

- hard code is there , this should be reduce

In [ ]: In [15]:

- many lines ====== single lines

- is your full run at any place?

- python

- data file

```
file_path="C:\\Users\\omkar\\OneDrive\\Documents\\Data science\\Naresh IT\\m
file=open(file_path,encoding='utf-8-sig')
file.read()
```

Out[15]: 'From stephen.marquard@uct.ac.za Sat Jan 5 09:14:16 2008\nReturn-Path:

<postmaster@collab.sakaiproject.org>\nReceived: from murder (mail.umich.edu [141.211.14.90])\n\t by frankenstein.mail.umich.edu (Cyrus v2.3.8) with LMTPA;\n\t Sat, 05 Jan 2008 09:14:16 -0500\nX-Sieve: CMU Sieve 2.3\nReceived: from murder ([unix socket])\n\t by mail.umich.edu (Cyrus v2.2.12) with LMTPA;\n\t Sat, 05 Jan 2008 09:14:16 -0500\nReceived: from holes.mr.itd.umich.edu (holes.mr.itd.umich.edu [141.211.14.79])\n\tby flawless.mail.umich.edu () with ESMTP id m05EEFR1013674;\n\tSat, 5 Jan 2008 09:14:15 -0500\nReceived: FROM paploo.uhi.ac.uk (app1.prod.collab.uhi.ac.uk [194.35.219.184])\n\tBY holes.mr.itd.umich.edu ID 477F90B0.2DB2F.12494 ; \n\t 5 Jan 2008 09:14:10 -0500\nReceived: from paploo.uhi.ac.uk (localhost [127.0.0.1])\n\tby paploo.uhi.ac.uk (Postfix) with ESMTP id 5F919BC2F2;\n\tSat, 5 Jan 2008 14:10:05 +0000 (GMT)\nMessage-ID: <200801051412.m05ECIaH010327@nakamura.uits.iupui.edu>\nMime-Version: 1.0\nContent-Transfer-Encoding: 7bit\nReceived: from prod.collab.uhi.ac.uk ([194.35.219.182])\n by paploo.uhi.ac.uk (JAMES SMTP Server 2.1.3) with SM TP ID 899\n for <source@collab.sakaiproject.org>;\n Sa t, 5 Jan 2008 14:09:50 +0000 (GMT)\nReceived: from nakamura.uits.iupui.e du (nakamura.uits.iupui.edu [134.68.220.122])\n\tby shmi.uhi.ac.uk (Post

fix) with ESMTP id A215243002\n\tfor <source@collab sakaiproject org>; S

```
file.read(
)
```

In [18]:

Out[18]: ''

In [17]:

```
file
```

Out[17]: <_io.TextIOWrapper name='C:\\Users\\omkar\\OneDrive\\Documents\\Data science\\Naresh IT\\mbox-short1.txt' mode='r' encoding='utf-8-sig'>

In [21]: In [23]:

```
Jan 2008 09:12:18 -05 00
Date: Sat, 5 Jan 2008 09:12:18 -0500
X-Authentication-Warning:
nakamura.uits.iupui.edu: apache set sender
to stephen.marquard@uct.ac.za using -f
To: source@collab.sakaiproject.org
From: stephen.marquard@uct.ac.za
Subject: [sakai] svn commit: r39772 -
content/branches/sakai_2-5-x/conte
nt-impl/impl/src/java/org/sakaiproject/con
tent/impl
X-Content-Type-Outer-Envelope: text/plain;
charset=UTF-8
X-Content-Type-Message-Body: text/plain;
charset=UTF-8
```

In [ ]:

```
#file_path="mbox-short.txt"
#file=open(file_path,encoding='utf-8-sig')
#print(file.read())

file_path="mbox-short.txt"
with open(file_path) as file:
 data=file.read()


print(data)
+0000 (GMT)
Received: from nakamura.uits.iupui.edu
(localhost [127.0.0.1])
```

```
Content Type: text/plain; charset UTF 8

####################
M-1#######################################
```

```
by nakamura.uits.iupui.edu
(8.12.11.20060308/8.12.11) with ESMTP id
m05ECJVp010329
for <source@collab.sakaiproject.org>; Sat,
5 Jan 2008 09:12:19 - 0500
Received: (from apache@localhost)
by nakamura.uits.iupui.edu
(8.12.11.20060308/8.12.11/Submit) id
m05ECIaH010327
for source@collab.sakaiproject.org; Sat, 5
```

```
print(file.read())

###################
M-2#####################################
######## file_path="mbox-short.txt"
file=open(file_path,encoding='utf-8-sig')
print(file.read())

###################
M-3#####################################
#

file_path="mbox-short.txt"
with open(file_path) as file:
 data=file.read()
```

In [25]:

```
file_path="C:\\Users\\omkar\\OneDrive\\Doc
uments\\Data science\\Naresh IT\\D
file=open(file_path,encoding='utf-8-sig')
```

```
len(data)
```

Out[25]: 94626

In [ ]:

```
# I want to get
```

number of lines

In [26]: In [27]:

```
5

0
9
:
1
4
:
1
6

2
0
0
8
```

```
for i in data:
 print(i)
n
```

R

```
file_path="mbox-short.txt"
file=open(file_path,encoding='utf-8-sig')
for i in file.read():
 print(i)

# are you getting lines
# or
# charcters
```

F

```
r
o
m

s
t
e
p
h
e
n
.
m
a
r
q
u
a
```

In [28]:

```
file_path="mbox-short.txt"
file=open(file_path,encoding='utf-8-sig')
```

In [29]:

```
for i in file:
 print(i)
```

```
From stephen.marquard@uct.ac.za Sat Jan 5
09:14:16 2008
```

```
Return-Path:
<postmaster@collab.sakaiproject.org>
```

```
Received: from murder (mail.umich.edu
[141.211.14.90])
```

```
 by frankenstein.mail.umich.edu (Cyrus
```

```
v2.3.8) with LMTPA;  Sat, 05 Jan 2008
```

```
09:14:16 -0500

X-Sieve: CMU Sieve 2.3

Received: from murder ([unix socket])

 by mail.umich.edu (Cyrus v2.2.12) with
LMTPA;

 Sat, 05 Jan 2008 09:14:16 -0500

Received: from holes.mr.itd.umich.edu
(holes.mr.itd.umich.edu [141.211.1
```

```
# extra space : 200 charcter

1910
In [33]: In [34]:
```

```
4 79])
```

before reading will have lines
if you want to see lines , work on open()
if you want to see the charcters , work on file.read()

```
#Q1) get number of lines

file_path="mbox-short.txt"
file=open(file_path,encoding='utf-8-sig')

count=0
for i in file:
 count+=1

print(count)

# I need to count how many lines
```

```
file_path="mbox-short.txt"
file=open(file_path,encoding='utf-8-sig')


for index,line in enumerate(file):
 print(index,line)
```

```
5 X-Sieve: CMU Sieve 2.3
```

```
6 Received: from murder ([unix socket])

7 by mail.umich.edu (Cyrus v2.2.12) with
LMTPA;

8 Sat, 05 Jan 2008 09:14:16 -0500

9 Received: from holes.mr.itd.umich.edu
(holes.mr.itd.umich.edu [141.21 1.14.79])
```

```
10 by flawless.mail.umich.edu () with

ESMTP id m05EEFR1013674; 11 Sat, 5 Jan

2008 09:14:15 -0500

12 Received: FROM paploo.uhi.ac.uk
(app1.prod.collab.uhi.ac.uk [194.35.2
19.184])
```

```
13 BY holes.mr.itd.umich.edu ID
477F90B0.2DB2F.12494 ;
```

```
# m-1:
file_path="mbox-short.txt"
file=open(file_path,encoding='utf-8-sig')


for index,line in enumerate(file):
 pass

number=index+1
print(number)
```

```
Out[36]: 1910
In [41]:
# Q3) print first 10 lines
file_path="mbox-short.txt"
file=open(file_path,encoding='utf-8-sig')

# for index,line in enumerate(file):
# print(index,line)
# if index==3:
```

```
1910


In [ ]: In [36]:


SyntaxError: (unicode error)
'unicodeescape' codec can't decode bytes
in pos


# m-2:
file_path="mbox-short.txt"
file=open(file_path,encoding='utf-8-sig')


l1=len([index for index,_ in
enumerate(file)])
l1

# here line are not required
# whenever you dont want use any specific
variable
# keep that variable name as _
```

```
# break

for index,line in enumerate(file):
 if index<10:
 print(index,line)

0 From stephen.marquard@uct.ac.za Sat Jan
```

```
5 09:14:16 2008 1 Return-Path:

<postmaster@collab.sakaiproject.org>

2 Received: from murder (mail.umich.edu

[141.211.14.90]) 3 by

frankenstein.mail.umich.edu (Cyrus

v2.3.8) with LMTPA; 4 Sat, 05 Jan 2008

09:14:16 -0500

5 X-Sieve: CMU Sieve 2.3

6 Received: from murder ([unix socket])

7 by mail.umich.edu (Cyrus v2.2.12) with
LMTPA;

8 Sat, 05 Jan 2008 09:14:16 -0500

9 Received: from holes.mr.itd.umich.edu
(holes.mr.itd.umich.edu [141.211.1 4.79])
In [42]: In [44]:
```

```
 print(line.strip())

From stephen.marquard@uct.ac.za Sat Jan 5
09:14:16 2008
```

```
Return-Path:
<postmaster@collab.sakaiproject.org>
Received: from murder (mail.umich.edu
[141.211.14.90])
by frankenstein.mail.umich.edu (Cyrus
v2.3.8) with LMTPA; Sat, 05 Jan 2008
09:14:16 -0500
X-Sieve: CMU Sieve 2.3
Received: from murder ([unix socket])
by mail.umich.edu (Cyrus v2.2.12) with
LMTPA;
Sat, 05 Jan 2008 09:14:16 -0500
Received: from holes.mr.itd.umich.edu
(holes.mr.itd.umich.edu [141.211.1 4.79])
by flawless.mail.umich.edu () with ESMTP
id m05EEFR1013674; Sat, 5 Jan 2008
09:14:15 -0500
Received: FROM paploo.uhi.ac.uk
(app1.prod.collab.uhi.ac.uk [194.35.219.
184])
BY holes.mr.itd.umich.edu ID
477F90B0.2DB2F.12494 ;
5 Jan 2008 09:14:10 -0500
Received: from paploo.uhi.ac.uk
(localhost [127.0.0.1])
by paploo.uhi.ac.uk (Postfix) with ESMTP
id 5F919BC2F2;
```

```
# Q4) make all lines in same alignment
file_path="mbox-short.txt"
file=open(file_path,encoding='utf-8-sig')

for line in file:
```

```
S t 5 J 2008 14 10 05 +0000 (GMT)

#Q5) Extract line those who are staring
with 'From'
# and keep in a list
# and get the count of list

file_path="mbox-short.txt"
file=open(file_path,encoding='utf-8-sig')

# for line in file:
# if line.startswith('From'):
# print(line)

from_list=[line for line in file if
line.startswith('From')] len(from_list)
```

```
Out[44]: 54
In [47]:                    [line.rstrip('\n') for
# Q6) remove '\n' from      line in from_list]
above output

Out[47]: ['From stephen.marquard@uct.ac.za Sat Jan 5 09:14:16 2008',
          'From: stephen.marquard@uct.ac.za',
          'From louis@media.berkeley.edu Fri Jan 4 18:10:48 2008',
          'From: louis@media.berkeley.edu',
          'From zqian@umich.edu Fri Jan 4 16:10:39 2008',
          'From: zqian@umich.edu',
          'From rjlowe@iupui.edu Fri Jan 4 15:46:24 2008',
          'From: rjlowe@iupui.edu',
          'From zqian@umich.edu Fri Jan 4 15:03:18 2008',
          'From: zqian@umich.edu',
          'From rjlowe@iupui.edu Fri Jan 4 14:50:18 2008',
          'From: rjlowe@iupui.edu',
          'From cwen@iupui.edu Fri Jan 4 11:37:30 2008',
          'From: cwen@iupui.edu',
          'From cwen@iupui.edu Fri Jan 4 11:35:08 2008',
          'From: cwen@iupui.edu',
          'From gsilver@umich.edu Fri Jan 4 11:12:37 2008',
          'From: gsilver@umich.edu',
          'From gsilver@umich.edu Fri Jan 4 11:11:52 2008',
          'From: gsilver@umich.edu',
          'From zqian@umich.edu Fri Jan 4 11:11:03 2008',
          'From: zqian@umich.edu',
          'From gsilver@umich.edu Fri Jan 4 11:10:22 2008',
          'From: gsilver@umich.edu',
          'From wagnermr@iupui.edu Fri Jan 4 10:38:42 2008',
          'From: wagnermr@iupui.edu',
          'From zqian@umich.edu Fri Jan 4 10:17:43 2008',
```

```
 'From: zqian@umich.edu',
 'From antranig@caret.cam.ac.uk Fri Jan 4 10:04:14 2008',
 'From: antranig@caret.cam.ac.uk',
 'From gopal.ramasammycook@gmail.com Fri Jan 4 09:05:31 2008',
 'From: gopal.ramasammycook@gmail.com',
 'From david.horwitz@uct.ac.za Fri Jan 4 07:02:32 2008',
 'From: david.horwitz@uct.ac.za',
 'From david.horwitz@uct.ac.za Fri Jan 4 06:08:27 2008',
 'From: david.horwitz@uct.ac.za',
 'From david.horwitz@uct.ac.za Fri Jan 4 04:49:08 2008',
 'From: david.horwitz@uct.ac.za',
 'From david.horwitz@uct.ac.za Fri Jan 4 04:33:44 2008',
 'From: david.horwitz@uct.ac.za',
 'From stephen.marquard@uct.ac.za Fri Jan 4 04:07:34 2008',
 'From: stephen.marquard@uct.ac.za',
 'From louis@media.berkeley.edu Thu Jan 3 19:51:21 2008',
 'From: louis@media.berkeley.edu',
 'From louis@media.berkeley.edu Thu Jan 3 17:18:23 2008',
 'From: louis@media.berkeley.edu',
 'From ray@media.berkeley.edu Thu Jan 3 17:07:00 2008',
 'From: ray@media.berkeley.edu',
 'From cwen@iupui.edu Thu Jan 3 16:34:40 2008',
 'From: cwen@iupui.edu',
 'From cwen@iupui.edu Thu Jan 3 16:29:07 2008',
 'From: cwen@iupui.edu',
 'From cwen@iupui.edu Thu Jan 3 16:23:48 2008',
 'From: cwen@iupui.edu']
```

In [48]:
```
#Q7)

#Q5) Extract line those who are staring with 'From'

from_list=[line for line in file if line.startswith('From')]
from_list

# Q6) remove '\n'

from_list=[line.rstrip('\n') for line in from_list]
from_list

file_path="mbox-short.txt"
file=open(file_path,encoding='utf-8-sig')


from_list=[line.rstrip('\n') for line in file if line.startswith('From')]
from_list
```

Out[48]:
```
['From stephen.marquard@uct.ac.za Sat Jan 5 09:14:16 2008',
       'From: stephen.marquard@uct.ac.za',
 'From louis@media.berkeley.edu Fri Jan 4 18:10:48 2008',
 'From: louis@media.berkeley.edu',
 'From zqian@umich.edu Fri Jan 4 16:10:39 2008',
 'From: zqian@umich.edu',
 'From rjlowe@iupui.edu Fri Jan 4 15:46:24 2008',
 'From: rjlowe@iupui.edu',
 'From zqian@umich.edu Fri Jan 4 15:03:18 2008',
 'From: zqian@umich.edu',
 'From rjlowe@iupui.edu Fri Jan 4 14:50:18 2008',
 'From: rjlowe@iupui.edu',
 'From cwen@iupui.edu Fri Jan 4 11:37:30 2008',
```

```
 'From: cwen@iupui.edu',
 'From cwen@iupui.edu Fri Jan 4 11:35:08 2008',
 'From: cwen@iupui.edu',
 'From gsilver@umich.edu Fri Jan 4 11:12:37 2008',
 'From: gsilver@umich.edu',
 'From gsilver@umich.edu Fri Jan 4 11:11:52 2008',
 'From: gsilver@umich.edu',
 'From zqian@umich.edu Fri Jan 4 11:11:03 2008',
 'From: zqian@umich.edu',
 'From gsilver@umich.edu Fri Jan 4 11:10:22 2008',
 'From: gsilver@umich.edu',
 'From wagnermr@iupui.edu Fri Jan 4 10:38:42 2008',
 'From: wagnermr@iupui.edu',
 'From zqian@umich.edu Fri Jan 4 10:17:43 2008',
 'From: zqian@umich.edu',
 'From antranig@caret.cam.ac.uk Fri Jan 4 10:04:14 2008',
 'From: antranig@caret.cam.ac.uk',
 'From gopal.ramasammycook@gmail.com Fri Jan 4 09:05:31 2008',
 'From: gopal.ramasammycook@gmail.com',
 'From david.horwitz@uct.ac.za Fri Jan 4 07:02:32 2008',
 'From: david.horwitz@uct.ac.za',
 'From david.horwitz@uct.ac.za Fri Jan 4 06:08:27 2008',
 'From: david.horwitz@uct.ac.za',
 'From david.horwitz@uct.ac.za Fri Jan 4 04:49:08 2008',
 'From: david.horwitz@uct.ac.za',
 'From david.horwitz@uct.ac.za Fri Jan 4 04:33:44 2008',
 'From: david.horwitz@uct.ac.za',
 'From stephen.marquard@uct.ac.za Fri Jan 4 04:07:34 2008',
 'From: stephen.marquard@uct.ac.za',
 'From louis@media.berkeley.edu Thu Jan 3 19:51:21 2008',
 'From: louis@media.berkeley.edu',
 'From louis@media.berkeley.edu Thu Jan 3 17:18:23 2008',
 'From: louis@media.berkeley.edu',
 'From ray@media.berkeley.edu Thu Jan 3 17:07:00 2008',
 'From: ray@media.berkeley.edu',
 'From cwen@iupui.edu Thu Jan 3 16:34:40 2008',
 'From: cwen@iupui.edu',
 'From cwen@iupui.edu Thu Jan 3 16:29:07 2008',
 'From: cwen@iupui.edu',
 'From cwen@iupui.edu Thu Jan 3 16:23:48 2008',
 'From: cwen@iupui.edu']
```

In [50]:
```
# Q8)
# extract all emails from above output
# 54 lines are there
# Each line has one email
# From'' to ''

emails=[line.split()[1] for line in from_list]
emails

str1='From stephen.marquard@uct.ac.za Sat Jan 5 09:14:16 2008'
str1.split()[1]
```

Out[50]:
```
['stephen.marquard@uct.ac.za',
 'stephen.marquard@uct.ac.za',
 'louis@media.berkeley.edu',
 'louis@media.berkeley.edu',
 'zqian@umich.edu',
 'zqian@umich.edu',
```

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

In [51]:

```
str1='From                          - extract your output
stephen.marquard@uct.ac.za Sat Jan 5
09:14:16 2008' str1.split()[1]      - then generalise the logic
```

- first take a single line

Out[51]: 'stephen.marquard@uct.ac.za'

In [ ]:

In [62]: In [71]: In [ ]:

```
#Q9)

# in above email list , there are
repeated emails are there # create unique
email list

# unique vowels
# counter
# summ
# create an emapty string/list/dictionary


email_list=[]
for email in emails:
 if email not in email_list:
 email_list.append(email)

print(len(email_list))
```

```
11
```

```
email_list=[]
[email_list.append(email) for email in
emails if email not in email_list]
print(email_list)
```

```
['stephen.marquard@uct.ac.za',
'louis@media.berkeley.edu',
'zqian@umich.ed u', 'rjlowe@iupui.edu',
'cwen@iupui.edu', 'gsilver@umich.edu',
'wagnermr@i upui.edu',
'antranig@caret.cam.ac.uk',
'gopal.ramasammycook@gmail.com', 'd
avid.horwitz@uct.ac.za',
'ray@media.berkeley.edu']
```

```
em
```