**TASK 4**

collect email

https://uetpeshawar.edu.pk/contactus.php

https://au.edu.pk/pages/About/Contact\_Us.aspx

http://www.uop.edu.pk/administration/

https://fia.gov.pk/contact

import re, urllib.request, time

emailRegex = re.compile(r'''

#example :

#something-.+\_@somedomain.com

(

([a-zA-Z0-9\_.+]+

@

[a-zA-Z0-9\_.+]+)

)

''', re.VERBOSE)

#Extacting Emails

def extractEmailsFromUrlText(urlText):

extractedEmail = emailRegex.findall(urlText)

allemails = []

for email in extractedEmail:

allemails.append(email[0])

lenh = len(allemails)

print("\tNumber of Emails : %s\n"%lenh )

seen = set()

for email in allemails:

if email not in seen: # faster than `word not in output`

seen.add(email)

emailFile.write(email+"\n")#appending Emails to a filerea

#HtmlPage Read Func

def htmlPageRead(url, i):

try:

start = time.time()

headers = { 'User-Agent' : 'Mozilla/5.0' }

request = urllib.request.Request(url, None, headers)

response = urllib.request.urlopen(request)

urlHtmlPageRead = response.read()

urlText = urlHtmlPageRead.decode()

print ("%s.%s\tFetched in : %s" % (i, url, (time.time() - start)))

extractEmailsFromUrlText(urlText)

except:

pass

#EmailsLeechFunction

def emailsLeechFunc(url, i):

try:

htmlPageRead(url,i)

except urllib.error.HTTPError as err:

if err.code == 404:

try:

url = 'http://webcache.googleusercontent.com/search?q=cache:'+url

htmlPageRead(url, i)

except:

pass

else:

pass

# TODO: Open a file for reading urls

start = time.time()

urlFile = open("/home/adikhan/Documents/codes/exam/urls.txt", 'r')

emailFile = open("emails.txt", 'a')

i=0

#Iterate Opened file for getting single url

for urlLink in urlFile.readlines():

urlLink = urlLink.strip('\'"')

i=i+1

emailsLeechFunc(urlLink, i)

print ("Elapsed Time: %s" % (time.time() - start))

urlFile.close()

emailFile.close()

print(EmailFile)

