Project 2: Automatic Data Carver using Python

Name: Shefali Athavale

Approach:

For this project, I wrote an Automatic Data Carver Program using Python. I used the following libraries for the same - sys, re, os, hashlib. The program asks the user to enter a file through the command line which is to be carved. It is assumed that the file will be a binary file. Two dictionaries are created for storing SOF and EOF signatures for JPEG/JPG, PNG and PDF files. Once the user enters the file, the program searches for Start of File (SOF) and End of File (EOF) offsets and stores them in lists. There are separate lists for SOF and EOF offsets according to the file types. Once we get the SOF and EOF offsets for the above file types, we carve the files and store it in the following format - 'carved_1.jpeg', 'carved_3.png', 'carved_5.pdf', etc. The program is tested on 'carve.lab' and 'midterm.dd' files. MD5 hash of all the carved files is also calculated to check if the files carved by the program match the files carved manually. The program outputs basic file information like the file name, file type, SOF and EOF offsets in bytes, File size and the MD5 hash of the file. The MD5 hashes are also stored in a file called 'hashes.txt'. All the carved files and the file 'hashes.txt' are stored in a directory titled by the last name('Athavale' in my case). Visual Studio Code is used for writing the program. The outputs from both the files and the code is attached below.

Output:

carve.lab:

Project2_ShefaliAthavale_Code.py

Project2_ShefaliAthavale_Code.py > ...

```
115         f.write("carved "+str(m)+" - "+hash)
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL

```
shefaliathavale@Shefalis-MacBook-Air Project 2 - Data Carver % /usr/bin/python3 "/Users/shefaliathavale/CUB_MS/Sem2/Digital Forensics/Project 2 -
 Data Carver/Project2_ShefaliAthavale_Code.py" carve.lab
Created file carved_1.jpeg
File Type is: jpeg
Start of File is: 43495424
End of File is: 43505613
File Size is: 10.189 KB
MD5 Hash of the file carved_1.jpeg is: b7e1fd222f355aeed11d3432c76fa2f7

Created file carved_2.jpeg
File Type is: jpeg
Start of File is: 143360
End of File is: 161589
File Size is: 18.229 KB
MD5 Hash of the file carved_2.jpeg is: 49c35f95eb60e987cb38e993b0deeeb3

Created file carved_3.png
File Type is: png
Start of File is: 163840
End of File is: 173135
File Size is: 9.295 KB
MD5 Hash of the file carved_3.png is: 29ce17a51d2bc8f56ff76e0aaf58c61e

Created file carved_4.png
File Type is: png
Start of File is: 42827776
End of File is: 43493582
File Size is: 665.806 KB
MD5 Hash of the file carved_4.png is: 14e82f5a7f8c8c06887d663cc3cc4945

Created file carved_5.pdf
File Type is: pdf
Start of File is: 4378624
End of File is: 4379102
File Size is: 0.478 KB
MD5 Hash of the file carved_5.pdf is: d28dff9e6ea67beb14b847ae42702f2e

Created file carved_6.pdf
File Type is: pdf
Start of File is: 4378624
End of File is: 4447020
File Size is: 68.396 KB
MD5 Hash of the file carved_6.pdf is: 9ecdba01deb179f49fd2fcc5f119a599

Created file carved_7.pdf
File Type is: pdf
Start of File is: 4378624
End of File is: 4379104
File Size is: 0.48 KB
MD5 Hash of the file carved_7.pdf is: 2b0518154d1d27d94ca2ae50d49d98fb
```

```
Created file carved_4.png
File Type is: png
Start of File is: 42827776
End of File is: 43493582
File Size is: 665.806 KB
MD5 Hash of the file carved_4.png is: 14e82f5a7f8c8c06887d663cc3cc4945

Created file carved_5.pdf
File Type is: pdf
Start of File is: 4378624
End of File is: 4379102
File Size is: 0.478 KB
MD5 Hash of the file carved_5.pdf is: d28dff9e6ea67beb14b847ae42702f2e

Created file carved_6.pdf
File Type is: pdf
Start of File is: 4378624
End of File is: 4447020
File Size is: 68.396 KB
MD5 Hash of the file carved_6.pdf is: 9ecdba01deb179f49fd2fcc5f119a599

Created file carved_7.pdf
File Type is: pdf
Start of File is: 4378624
End of File is: 4379104
File Size is: 0.48 KB
MD5 Hash of the file carved_7.pdf is: 2b0518154d1d27d94ca2ae50d49d98fb

Created file carved_8.pdf
File Type is: pdf
Start of File is: 4378624
End of File is: 4447022
File Size is: 68.398 KB
MD5 Hash of the file carved_8.pdf is: 02b5be6cd684487897ae852b744aaac2

shefaliathavale@Shefalis-MacBook-Air Project 2 - Data Carver %
```

Athavale

Name	Date Modified	Size	Kind
carved_1.jpeg	Today at 2:56 AM	10 KB	JPEG image
carved_2.jpeg	Today at 2:56 AM	18 KB	JPEG image
carved_3.png	Today at 2:56 AM	9 KB	PNG image
carved_4.png	Today at 2:56 AM	666 KB	PNG image
carved_5.pdf	Today at 2:56 AM	478 bytes	PDF Document
carved_6.pdf	Today at 2:56 AM	68 KB	PDF Document
carved_7.pdf	Today at 2:56 AM	480 bytes	PDF Document
carved_8.pdf	Today at 2:56 AM	68 KB	PDF Document
hashes.txt	Today at 2:56 AM	352 bytes	Plain Text

midterm.dd:

```
Project2_ShefaliAthavale_Code.py
Project2_ShefaliAthavale_Code.py > ...

PROBLEMS    OUTPUT    DEBUG CONSOLE    TERMINAL

shefaliathavale@Shefalis-MacBook-Air Project 2 - Data Carver % /usr/bin/python3 "/Users/shefaliathavale/CUB_MS/Sem2/Digital Forensics/Project 2 -
 Data Carver/Project2_ShefaliAthavale_Code.py" midterm.dd
Created file carved_1.jpeg
File Type is: jpeg
Start of File is: 1732608
End of File is: 1774488
File Size is: 41.88 KB
MD5 Hash of the file carved_1.jpeg is: b88fa6ddc8e5f5fa21dc7f13cfae46ec

Created file carved_2.jpeg
File Type is: jpeg
Start of File is: 1789952
End of File is: 1797363
File Size is: 7.411 KB
MD5 Hash of the file carved_2.jpeg is: acb990ea4affbed8f40c77e3236efeba

Created file carved_3.png
File Type is: png
Start of File is: 1409024
End of File is: 1512621
File Size is: 103.597 KB
MD5 Hash of the file carved_3.png is: 5ebff07cd1965cfcfbaf34a2f4f99518

Created file carved_4.png
File Type is: png
Start of File is: 1781760
End of File is: 1783700
File Size is: 1.94 KB
MD5 Hash of the file carved_4.png is: 0adba700e54834d1aa6eb91e5a7bdcb2

Created file carved_5.png
File Type is: png
Start of File is: 1814528
End of File is: 1842941
File Size is: 28.413 KB
MD5 Hash of the file carved_5.png is: 4f8e8263c50341f3966e759eb4865027

Created file carved_6.png
File Type is: png
Start of File is: 1843200
End of File is: 1871613
File Size is: 28.413 KB
MD5 Hash of the file carved_6.png is: 4f8e8263c50341f3966e759eb4865027
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL

```
Created file carved_6.png
File Type is: png
Start of File is: 1843200
End of File is: 1871613
File Size is: 28.413 KB
MD5 Hash of the file carved_6.png is: 4f8e8263c50341f3966e759eb4865027

Created file carved_7.pdf
File Type is: pdf
Start of File is: 1048576
End of File is: 1057444
File Size is: 8.868 KB
MD5 Hash of the file carved_7.pdf is: ef79d46b5cb0d39189cf1f87b29b72eb

Created file carved_8.pdf
File Type is: pdf
Start of File is: 1163264
End of File is: 1172134
File Size is: 8.87 KB
MD5 Hash of the file carved_8.pdf is: eeb97a8afad7dd7d4d302094a324fca7

Created file carved_9.pdf
File Type is: pdf
Start of File is: 1347584
End of File is: 1356467
File Size is: 8.883 KB
MD5 Hash of the file carved_9.pdf is: 4e528fa475d81c653bd920110fe95a3b

Created file carved_10.pdf
File Type is: pdf
Start of File is: 1359872
End of File is: 1368745
File Size is: 8.873 KB
MD5 Hash of the file carved_10.pdf is: 44d1a898e20946e0f309f916ffe2262b

Created file carved_11.pdf
File Type is: pdf
Start of File is: 1531904
End of File is: 1545022
File Size is: 13.118 KB
MD5 Hash of the file carved_11.pdf is: e1e49fe1208dfa16b2534a4a2a9637d4

Created file carved_12.pdf
File Type is: pdf
Start of File is: 1556480
End of File is: 1568274
File Size is: 11.794 KB
MD5 Hash of the file carved_12.pdf is: 4e66a7a01f285b4b45293c08bcb95862
```

```
Created file carved_12.pdf
File Type is: pdf
Start of File is: 1556480
End of File is: 1568274
File Size is: 11.794 KB
MD5 Hash of the file carved_12.pdf is: 4e66a7a01f285b4b45293c08bcb95862

Created file carved_13.pdf
File Type is: pdf
Start of File is: 1568768
End of File is: 1577632
File Size is: 8.864 KB
MD5 Hash of the file carved_13.pdf is: 620971e059ae13a9292e74091196d7af

Created file carved_14.pdf
File Type is: pdf
Start of File is: 1617920
End of File is: 1632181
File Size is: 14.261 KB
MD5 Hash of the file carved_14.pdf is: 9a56f91220a636cea7f022f0d7f6c298

Created file carved_15.pdf
File Type is: pdf
Start of File is: 1720320
End of File is: 1729196
File Size is: 8.876 KB
MD5 Hash of the file carved_15.pdf is: 12745d39a653f280fff8d454732e0bae

shefaliathavale@Shefalis-MacBook-Air Project 2 - Data Carver %
```

Athavale

Name	Date Modified	Size	Kind
carved_1.jpeg	Today at 2:59 AM	42 KB	JPEG image
carved_2.jpeg	Today at 2:59 AM	7 KB	JPEG image
carved_3.png	Today at 2:59 AM	104 KB	PNG image
carved_4.png	Today at 2:59 AM	2 KB	PNG image
carved_5.pdf	Today at 2:56 AM	478 bytes	PDF Document
carved_5.png	Today at 2:59 AM	28 KB	PNG image
carved_6.pdf	Today at 2:56 AM	68 KB	PDF Document
carved_6.png	Today at 2:59 AM	28 KB	PNG image
carved_7.pdf	Today at 2:59 AM	9 KB	PDF Document
carved_8.pdf	Today at 2:59 AM	9 KB	PDF Document
carved_9.pdf	Today at 2:59 AM	9 KB	PDF Document
carved_10.pdf	Today at 2:59 AM	9 KB	PDF Document
carved_11.pdf	Today at 2:59 AM	13 KB	PDF Document
carved_12.pdf	Today at 2:59 AM	12 KB	PDF Document
carved_13.pdf	Today at 2:59 AM	9 KB	PDF Document
carved_14.pdf	Today at 2:59 AM	14 KB	PDF Document
carved_15.pdf	Today at 2:59 AM	9 KB	PDF Document
hashes.txt	Today at 2:59 AM	666 bytes	Plain Text

Code:

```
## File Signatures of PNG, JPEG/JPG and PDF files

# PNG: 89 50 4E 47 0D 0A 1A 0A - 49 45 4E 44 AE 42 60 82

# JPEG/JPG: FF D8 - FF D9

# PDF: 25 50 44 46 -
# 0A 25 25 45 4F 46 (.%%EOF)
# 0A 25 25 45 4F 46 0A (.%%EOF.)
# 0D 0A 25 25 45 4F 46 0D 0A (..%%EOF..)
# 0D 25 25 45 4F 46 0D (.%%EOF.)

import sys
import re
import os
import hashlib

md5Hashes = []

os.makedirs('Athavale',exist_ok=True)
## Function to carve file and save individual files
def carveFile(sof,eof,subdata,c,type):
   fileName = 'carved_'+str(c)+'.'+str(type)
   fcarve = open(fileName, 'wb')
   fcarve.write(subdata)
   fcarve.close()
   print("Created file "+fileName)
   print("File Type is: "+type)
   print("Start of File is: "+str(sof))
   print("End of File is: "+str(eof))
   print("File Size is: "+str((eof-sof)/1000)+" KB")
   with open(fileName,'rb') as f:
       data = f.read()
       md5Hash = hashlib.md5(data).hexdigest()
       print("MD5 Hash of the file "+fileName+ " is: "+str(md5Hash))
       md5Hashes.append(md5Hash)
   print()
```

```
## Start of File and End of File for JPEG/JPG, PNG, PDF stored in dictionary
sof_dict =
{'jpeg':b'\xFF\xD8\xFF\xE0','jpeg1':b'\xFF\xD8\xFF\xE1','png':b'\x89\x50\x4E\x47\x0D\x
0A\x1A\x0A','pdf':b'\x25\x50\x44\x46'}
eof_dict =
{'jpeg':b'\xFF\xD9\x00','png':b'\x49\x45\x4E\x44\xAE\x42\x60\x82','pdf':b'\x0A\x25\x25
\x45\x4F\x46','pdf1':b'\x0A\x25\x25\x45\x4F\x46\x0A','pdf2':b'\x0D\x0A\x25\x25\x45\x4F
\x46\x0D\x0A', 'pdf3':b'\x0D\x25\x25\x45\x4F\x46\x0D'}

## File name of the file to be carved
fname = str(sys.argv[1])
fname_obj = open(fname, 'rb')
data = fname_obj.read()
fname_obj.close()

## List of Start of File and End of File offsets for JPEG/JPG Images
sof_list_jpeg=[match.start() for match in
re.finditer(re.escape(sof_dict['jpeg']),data)]
sof_list_jpeg.extend(match.start() for match in
re.finditer(re.escape(sof_dict['jpeg1']),data))
eof_list_jpeg=[match.start()+2 for match in
re.finditer(re.escape(eof_dict['jpeg']),data)]

## List of Start of File and End of File offsets for PNG Images
sof_list_png=[match.start() for match in re.finditer(re.escape(sof_dict['png']),data)]
eof_list_png=[match.start()+8 for match in
re.finditer(re.escape(eof_dict['png']),data)]

## List of Start of File and End of File offsets for PDFs
sof_list_pdf=[match.start() for match in re.finditer(re.escape(sof_dict['pdf']),data)]
eof_list_pdf=[match.start()+6 for match in
re.finditer(re.escape(eof_dict['pdf']),data)]
eof_list_pdf.extend(match.start()+7 for match in
re.finditer(re.escape(eof_dict['pdf1']),data))
eof_list_pdf.extend(match.start()+9 for match in
re.finditer(re.escape(eof_dict['pdf2']),data))
eof_list_pdf.extend(match.start()+7 for match in
re.finditer(re.escape(eof_dict['pdf3']),data))

c = 0
os.chdir('Athavale')
```

```
## Loop to get the JPEG/JPG data from SOF and EOF offsets and send to carveFile
function to carve the file
for i in sof_list_jpeg:
   flag = 0
   for j in eof_list_jpeg:
       if i<j and flag==0:
           flag = 1

           subdata = data[i:j]
           c+=1
           carveFile(i,j,subdata,c,'jpeg')

## Loop to get the PNG data from SOF and EOF offsets and send to carveFile function to
carve the file
for i in sof_list_png:
   flag = 0
   for j in eof_list_png:
       if i<j and flag==0:
           flag = 1
           subdata = data[i:j]
           c+=1
           carveFile(i,j,subdata,c,'png')

## Loop to get the PDF data from SOF and EOF offsets and send to carveFile function to
carve the file
for i in sof_list_pdf:
   flag = 0
   for j in eof_list_pdf:
       if i<j and flag==0:
           flag = 1
           subdata = data[i:j]
           c+=1
           carveFile(i,j,subdata,c,'pdf')

m = 1
with open('hashes.txt','w') as f:
   for hash in md5Hashes:
       f.write("carved_"+str(m)+" - "+hash)
       f.write("\n")
       m+=1
   f.close()
sys.exit()
```