Michael Steven Towns

The first thing I did was to download the corrupted.docx, after it failed to open i looked up base64 decoding in python, leading me to the base64 module and the command “base64.b64decode(inFile.read())” which returns a decoded string. I saved this string to a text file for viewing. After this I opened it up in vim which allows me to see non-ascii characters as hex values. Opening it up I could see “%PDF” near the top, i then looked up magic numbers on wikipedia leading me to the page on magic numbers and the list of file signatures. After attempting to read from the start of one file to the start of the next unsuccessfully, I realized the files must be padded with extra characters, so I looked up the individual formats on wikipedia where in the body of the text they give the end code for most of the formats I needed. From this I was able to pull out the JPEG, PDF, and PNG. the GIF wikipedia page did not list an explicit end for the file so after I had all the others, i simply worked my way backwards from the PNG, removing a single character and then testing if that was the correct file, it worked after 16 characters.

To find the 5th file I looked at the file from the end of the PNG, and then looked through the wikipedia page for list of file signatures, i looked in the major file formats that I thought it might be: ZIP, DOC, DOCX, RAR, and others that are fairly common. I found PK it in the DOCX/ZIP section and the file had a PK near the beginning. So I assumed PK was the start and from there I worked backwards from the end of the file removing a single character and opening it in libreoffice until i got the document.

Relevant pages:
https://en.wikipedia.org/wiki/Magic_number_%28programming%29
https://en.wikipedia.org/wiki/List_of_file_signatures

https://en.wikipedia.org/wiki/JPEG
https://en.wikipedia.org/wiki/Portable_Document_Format
https://en.wikipedia.org/wiki/GIF
https://en.wikipedia.org/wiki/Portable_Network_Graphics
https://en.wikipedia.org/wiki/Zip_%28file_format%29

How To Run:
Requires Python 2
Input file:
    corrupted.docx
Output files:
    f1.jpeg
    f2.pdf
    f3.gif
    f4.png
    f5.docx
Run using:
    python2 file_carving.py
It will open corrupted.docx and then create each of the output files, no user input