OSINT and Forensic Challenge from nw3c

By Hugh Briggs

Summary: This challenge involved parsing out Instagram info from a MIME HTML file to help answer three questions. In this writeup, I will walk through multiple Techniques, Tools, Procedures (TTP's) to get the various flags.

Summary of Questions:

- What is the URL of the saved webpage?
- When did Babacanbabajo comment on the post?
  - Put in 24hr UTC Flag Format: yyyy-mm-dd hh:mm:ss
- There is a picture inside the file that Instagram has embedded. It has two words approximately in the middle surrounded by other images/icons. They both start with
  the letter F. What are the two words in order starting with the one on top?

What is MHTML?

"MHTML, an initialism of MIME encapsulation of aggregate HTML documents, is a web page archive format used to combine, in a single computer file, the HTML code and its companion resources (such as images, Flash animations, Java applets, and audio and video files) that are represented by external hyperlinks in the web page's HTML code. The content of an MHTML file is encoded using the same techniques that were first developed for HTML email messages, using the MIME content type multipart/related.[1] MHTML files use a .mhtml or .mht filename extension."

-- from https://en.wikipedia.org/wiki/MHTML

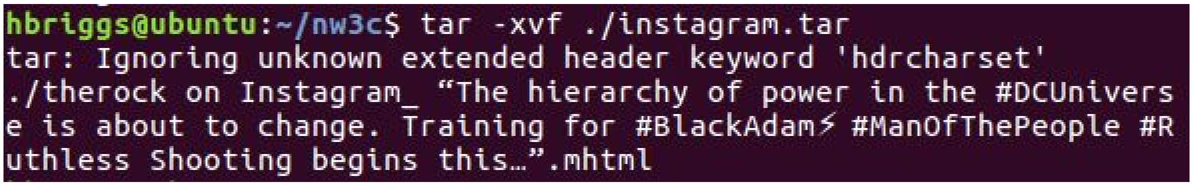
More info on:

- Nw3c - host of CTF https://nw3.ctfd.io
- MIME HTML https://tools.ietf.org/html/rfc2557
- Instagram Developer resource
  https://developers.facebook.com/docs/instagram

Below are details from the challenge that apply to all the questions:

- Download and use Instagram Post.tar file and use it to answer the three questions for today.
- Instagram Post.tar MD5=a8858d76e7eb56f698e1cacbee05ef65

Extracting the tar file will produce the following in a linux terminal. This file will be the base for answering the questions.



```
hbriggs@ubuntu:~/nw3c$ tar -xvf ./instagram.tar
tar: Ignoring unknown extended header keyword 'hdrcharset'
./therock on Instagram_ “The hierarchy of power in the #DCUnivers
e is about to change. Training for #BlackAdam⚡ #ManOfThePeople #R
uthless Shooting begins this…”.mhtml
```

Let’s get started with the first question.

1st Question - The Rock

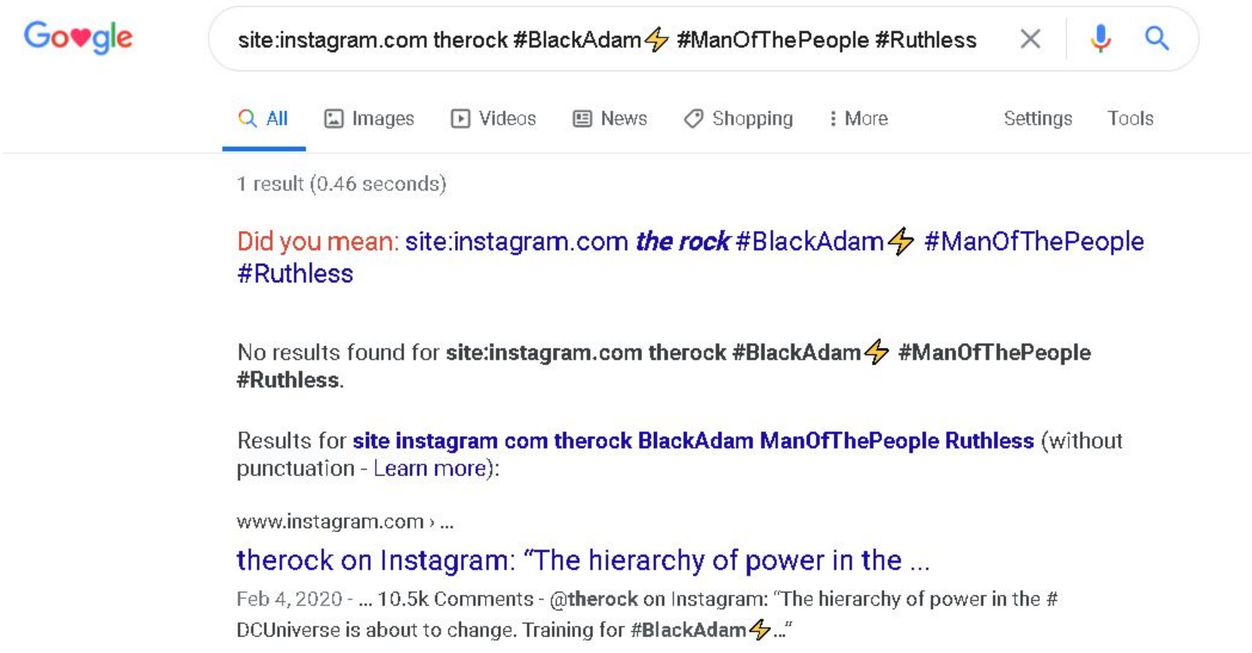
- What is the URL of the saved webpage?

The string in the mhtml: therock on Instagram_ “The hierarchy of power in the #DCUniverse is about to change. Training for #BlackAdam⚡ #ManOfThePeople #Ruthless Shooting begins this…”.mhtml

1st TTP - Google it!

We know that from the string, it involves the Rock, instagram, and some hashtags. These should be enough to help narrow down the millions of posts in Instagram. Lets limit results to only the instagram site. The below query will do the trick.

- site:instagram.com therock #BlackAdam⚡ #ManOfThePeople #Ruthless


Google
site:instagram.com therock #BlackAdam⚡ #ManOfThePeople #Ruthless
All
Images
Videos
News
Shopping
More
Settings
Tools

1 result (0.46 seconds)

Did you mean: site:instagram.com ***the rock*** #BlackAdam⚡ #ManOfThePeople #Ruthless

No results found for **site:instagram.com therock #BlackAdam⚡ #ManOfThePeople #Ruthless**.

Results for **site instagram com therock BlackAdam ManOfThePeople Ruthless** (without punctuation - Learn more):

www.instagram.com › ...

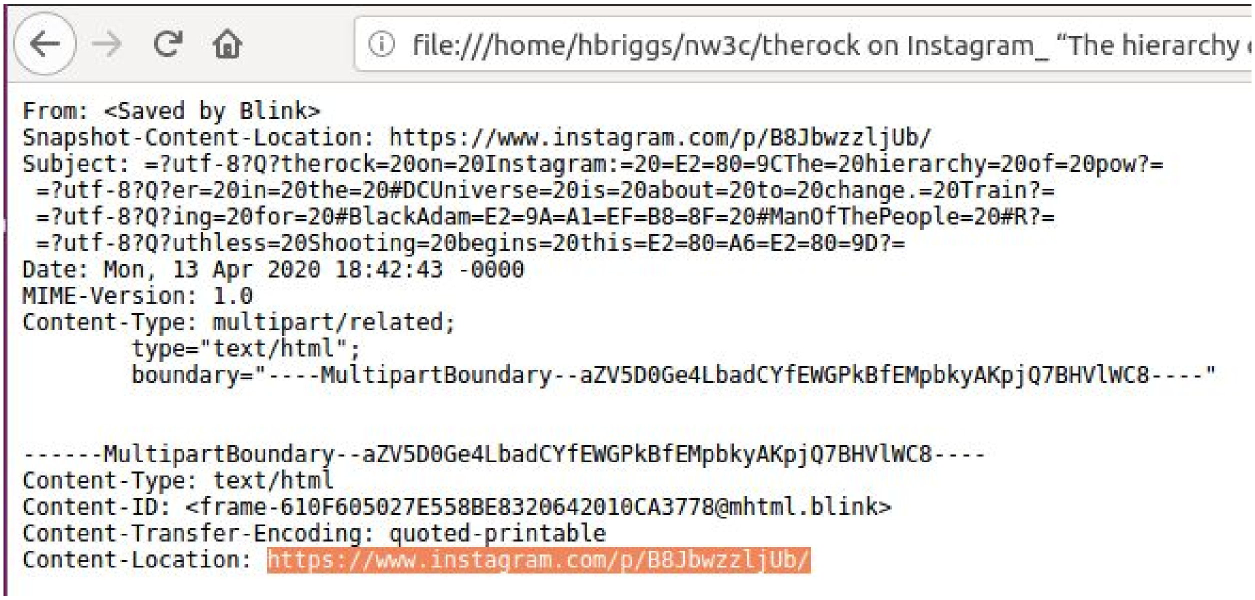
therock on Instagram: "The hierarchy of power in the ...

Feb 4, 2020 - ... 10.5k Comments - @**therock** on Instagram: "The hierarchy of power in the # DCUniverse is about to change. Training for #**BlackAdam**⚡..."

You can then copy the link address or click on the link to see the URL in the browser.

2nd TTP - Search the mhtml file

An mhtml file is text that contains html code, so you can open it up in a text editor or browser. Opening it up in firefox, we can see the URL is at the beginning of the file. The MultipartBoundary is part of the MIME structure to help organize the file.


file:///home/hbriggs/nw3c/therock on Instagram_ "The hierarchy

```
From: <Saved by Blink>
Snapshot-Content-Location: https://www.instagram.com/p/B8JbwzzljUb/
Subject: =?utf-8?Q?therock=20on=20Instagram:=20=E2=80=9CThe=20hierarchy=20of=20pow?=
 =?utf-8?Q?er=20in=20the=20#DCUniverse=20is=20about=20to=20change.=20Train?=
 =?utf-8?Q?ing=20for=20#BlackAdam=E2=9A=A1=EF=B8=8F=20#ManOfThePeople=20#R?=
 =?utf-8?Q?uthless=20Shooting=20begins=20this=E2=80=A6=E2=80=9D?=
Date: Mon, 13 Apr 2020 18:42:43 -0000
MIME-Version: 1.0
Content-Type: multipart/related;
        type="text/html";
        boundary="----MultipartBoundary--aZV5D0Ge4LbadCYfEWGPkBfEMpbkyAKpjQ7BHVlWC8----"


------MultipartBoundary--aZV5D0Ge4LbadCYfEWGPkBfEMpbkyAKpjQ7BHVlWC8----
Content-Type: text/html
Content-ID: <frame-610F605027E558BE8320642010CA3778@mhtml.blink>
Content-Transfer-Encoding: quoted-printable
Content-Location: https://www.instagram.com/p/B8JbwzzljUb/
```

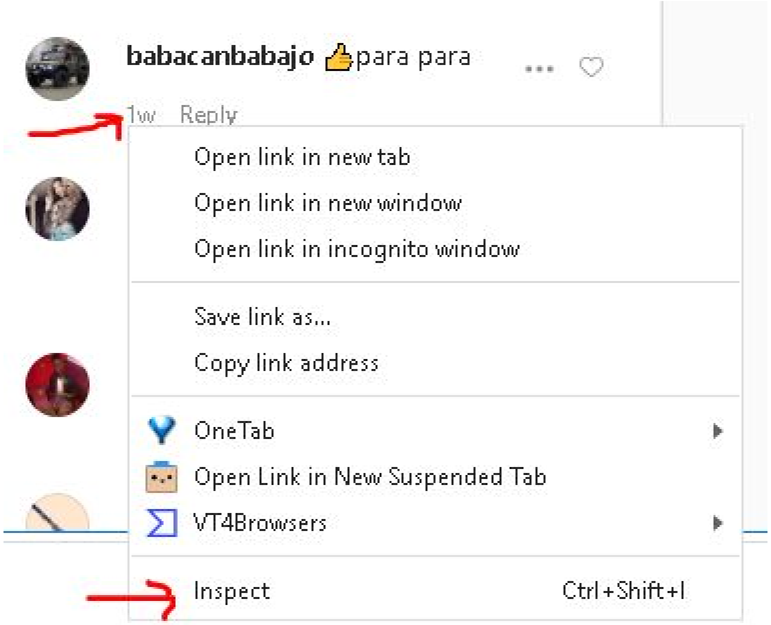
FLAG: https://www.instagram.com/p/B8JbwzzljUb/

2nd question - Babacanbabajo

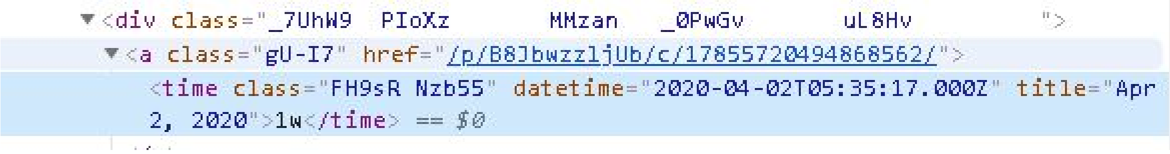
- When did Babacanbabajo comment on the post?
  - Put in 24hr UTC Flag Format: yyyy-mm-dd hh:mm:ss

1st TTP - Use the browser, Luke

From the question, we should look for that user Babacanbabajo. Luckily, that user appeared right away in the post. Right click on the time and then click on “Inspect”.



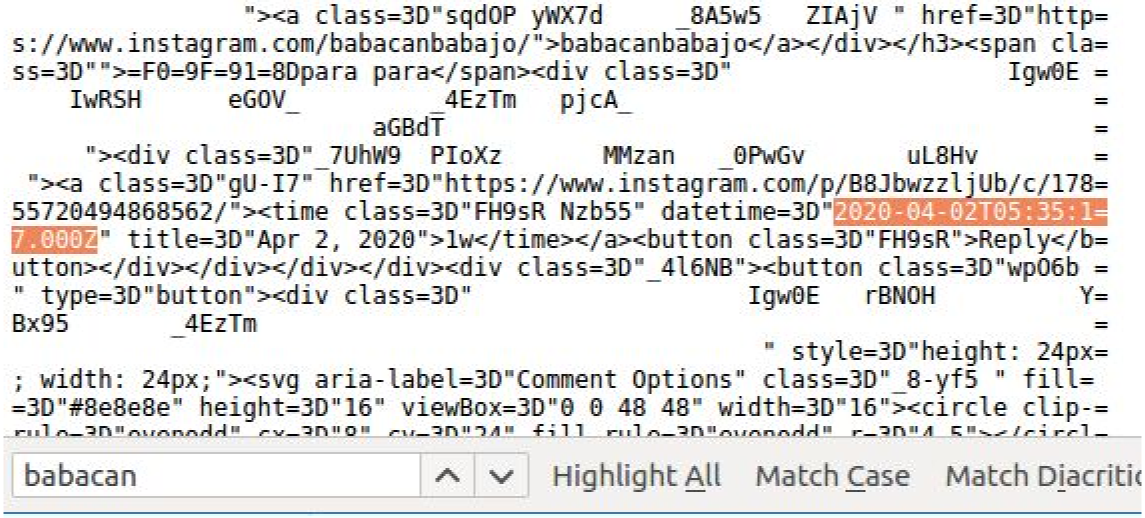
You will then be in the developer’s console of the web browser and it should drop you into where the user’s timestamp is.



```
▼<div class="_7UhW9  PIoXz       MMzan   _0PwGv       uL8Hv          ">
  ▼<a class="gU-I7" href="/p/B8JbwzzljUb/c/17855720494868562/">
     <time class="FH9sR Nzb55" datetime="2020-04-02T05:35:17.000Z" title="Apr
     2, 2020">1w</time> == $0
```

2nd TTP - Ctrl+F is your friend

Within the file, we can use “Find” on the string “Babacanbabajo” and grab the time, though it is a little harder to read.



```
                 "><a class=3D"sqdOP yWX7d     _8A5w5   ZIAjV " href=3D"http=
s://www.instagram.com/babacanbabajo/">babacanbabajo</a></div></h3><span cla=
ss=3D"">=F0=9F=91=8Dpara para</span><div class=3D"                  Igw0E =
    IwRSH      eGOV_         _4EzTm   pjcA_                              =
                        aGBdT                                            =
     "><div class=3D"_7UhW9  PIoXz       MMzan    _0PwGv       uL8Hv       =
 "><a class=3D"gU-I7" href=3D"https://www.instagram.com/p/B8JbwzzljUb/c/178=
55720494868562/"><time class=3D"FH9sR Nzb55" datetime=3D"2020-04-02T05:35:1=
7.000Z" title=3D"Apr 2, 2020">1w</time></a><button class=3D"FH9sR">Reply</b=
utton></div></div></div></div><div class=3D"_4l6NB"><button class=3D"wpO6b =
" type=3D"button"><div class=3D"                     Igw0E   rBNOH         Y=
Bx95       _4EzTm                                                        =
                                                    " style=3D"height: 24px=
; width: 24px;"><svg aria-label=3D"Comment Options" class=3D"_8-yf5 " fill=
=3D"#8e8e8e" height=3D"16" viewBox=3D"0 0 48 48" width=3D"16"><circle clip-=
rule=3D"evenodd" cx=3D"8" cy=3D"24" fill-rule=3D"evenodd" r=3D"4.5"></circl=
```

babacan

2020-04-02T05:35:17.000Z can then be converted to 2020-04-02 05:35:17 so that it matches the flag format.

FLAG: 2020-04-02 05:35:17

3rd Question - Added Data

- There is a picture inside the file that Instagram has embedded. It has two words approximately in the middle surrounded by other images/icons. They both start with the letter F. What are the two words in order starting with the one on top?

Mhtml files are nothing but HTML code, so how can we find an image?

Luckily, we can search for images in multiple ways in HTML. We know that images in HTML usually have the tag “<img” followed by something else. But this did not yield results. Searching by .jpg also did not help, however you can pull up personal images related to the post. This may be useful for other OSINT reconnaissance, such as grabbing only .jpg in the mhtml file and mapping to users.

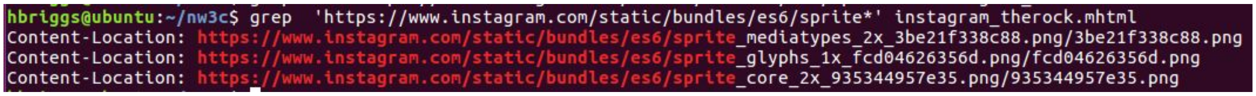
When I searched for .png, there were 59 matches, but I did notice a trend that it was pulling up results that had the /static/bundles/ directory as part of the link. Another iteration through these results revealed a /static/bundles/es6/ directory.

In the 3rd question there was this text: “in the middle surrounded by other images/icons”. Hint* This line is referring to image sprites. An image “sprite” can be defined as a collection of icons and images all inside one image. (from https://www.w3schools.com/css/css_image_sprites.asp)

The image with the flag should have a collection of icons and images all inside it and this image should be an image sprite.

Now Instagram does have a fair number of icons, so it would be safe to assume they also use image sprites. With the knowledge of image sprite and .png, was able to grep out directories that had the word sprite in them, narrowing it down to three .png files. (I did try combinations of image and sprite, but sprite was all that was needed.)

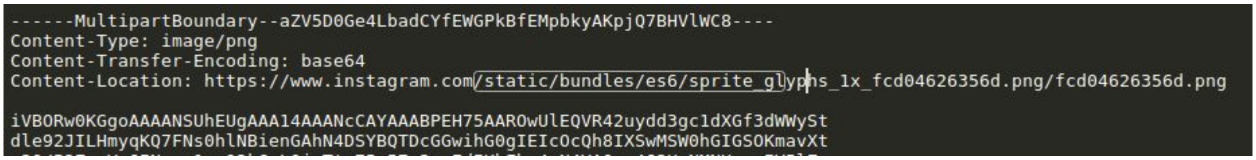
Going through the three .png files, fcd04626356d.png was the one with the flag. But how to show the image?



```
hbriggs@ubuntu:~/nw3c$ grep  'https://www.instagram.com/static/bundles/es6/sprite*' instagram_therock.mhtml
Content-Location: https://www.instagram.com/static/bundles/es6/sprite_mediatypes_2x_3be21f338c88.png/3be21f338c88.png
Content-Location: https://www.instagram.com/static/bundles/es6/sprite_glyphs_1x_fcd04626356d.png/fcd04626356d.png
Content-Location: https://www.instagram.com/static/bundles/es6/sprite_core_2x_935344957e35.png/935344957e35.png
```

TTP 1 - Chop it up with Cyberchef

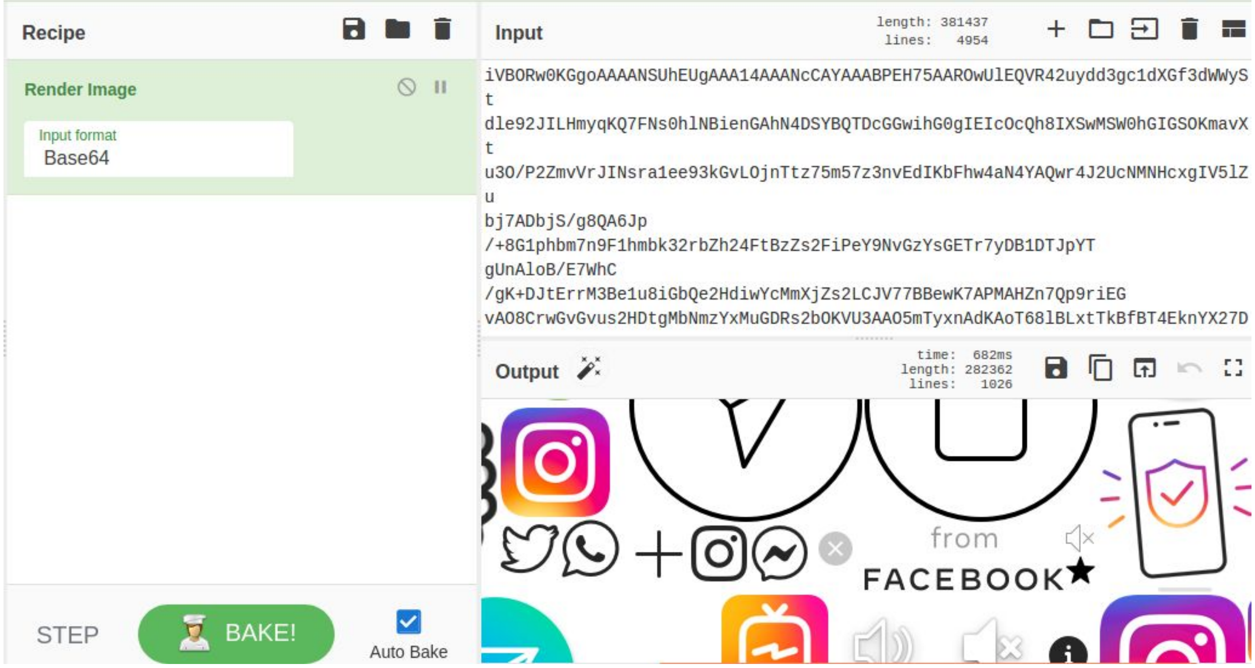
We know that the multipart boundary helps organize the html file and gives details on objects, even images. Let’s look for the .png file and MultipartBoundary in the mhtml.



```
------MultipartBoundary--aZV5D0Ge4LbadCYfEWGPkBfEMpbkyAKpjQ7BHVlWC8----
Content-Type: image/png
Content-Transfer-Encoding: base64
Content-Location: https://www.instagram.com/static/bundles/es6/sprite_glyphs_1x_fcd04626356d.png/fcd04626356d.png

iVBORw0KGgoAAAANSUhEUgAAA14AAANcCAYAAABPEH75AAROwUlEQVR42uydd3gc1dXGf3dWWySt
dle92JILHmyqKQ7FNs0hlNBienGAhN4DSYBQTDcGGwihG0gIEIcOcQh8IXSwMSW0hGIGSOKmavXt
```

The key thing to notice here is that the .png is base64 encoded. So that big block of text below Content-Location is the base64 encoding of the image. We can copy the whole block of text into cyberchef and use “Render Image” to output an image with the flag.



Input format
Base64

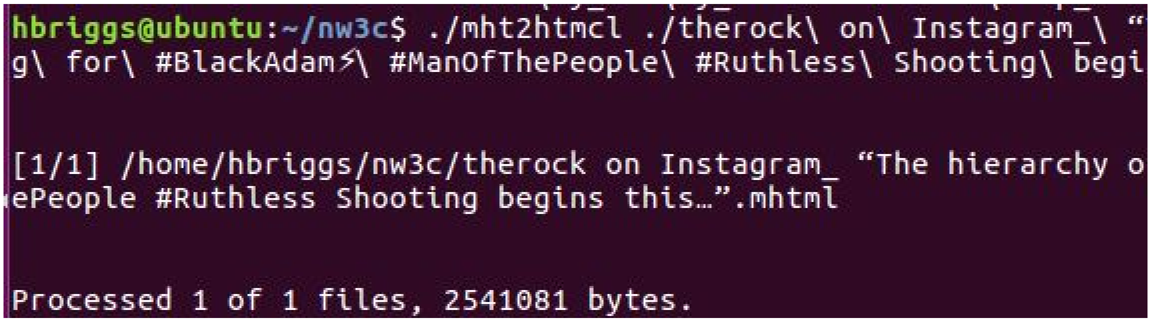
length: 381437
lines: 4954

iVBORw0KGgoAAAANSUhEUgAAA14AAANcCAYAAABPEH75AAROwUlEQVR42uydd3gc1dXGf3dWWyS
t
dle92JILHmyqKQ7FNs0hlNBienGAhN4DSYBQTDcGGwihG0gIEIcOcQh8IXSwMSW0hGIGSOKmavX
t
u3O/P2ZmvVrJINsra1ee93kGvLOjnTtz75m57z3nvEdIKbFhw4aN4YAQwr4J2UcNMNHcxgIV5lZ
u
bj7ADbjS/g8QA6Jp
/+8G1phbm7n9F1hmbk32rbZh24FtBzZs2FiPeY9NvGzYsGETr7yDB1DTJpYT
gUnAloB/E7WhC
/gK+DJtErrM3Be1u8iGbQe2HdiwYcMmXjZs2LCJV77BBewK7APMAHZn7Qp9riEG
vAO8CrwGvGvus2HDtgMbNmzYxMuGDRs2bOKVU3AA05mTyxnAdKAoT68lBkfBfBT4EknYX27D

time: 682ms
length: 282362
lines: 1026

TTP 2 - mht2htm Mutation

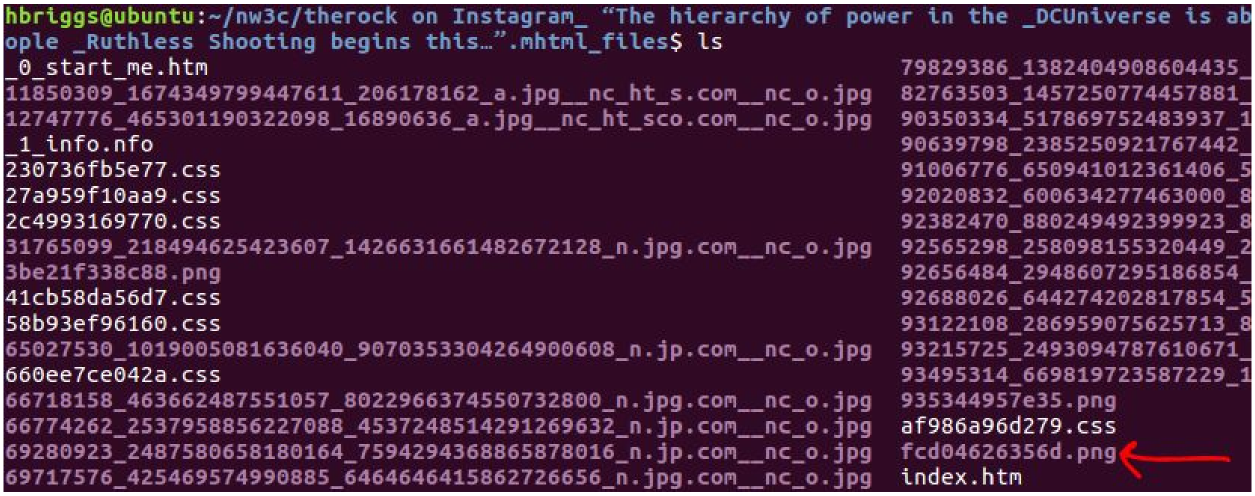
https://sourceforge.net/projects/mht2htm/
This program will take the mhtml file and convert it to html and put everything in a folder, making it nice and clean. You can go through each picture and find the flag.



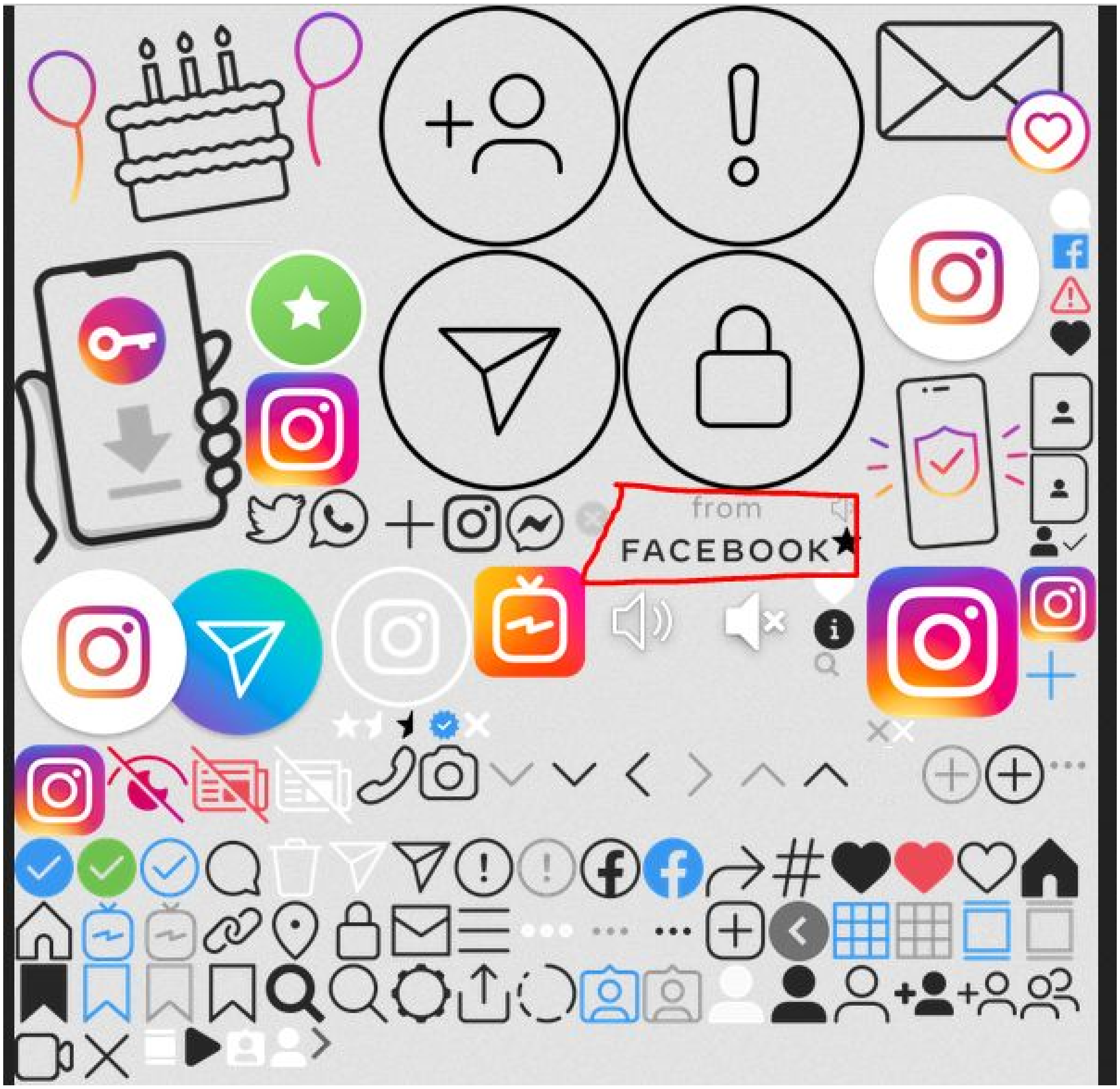
```
hbriggs@ubuntu:~/nw3c$ ./mht2htmcl ./therock\ on\ Instagram_\ “
g\ for\ #BlackAdam⚡\ #ManOfThePeople\ #Ruthless\ Shooting\ begi


[1/1] /home/hbriggs/nw3c/therock on Instagram_ “The hierarchy o
ePeople #Ruthless Shooting begins this…”.mhtml


Processed 1 of 1 files, 2541081 bytes.
```



```
hbriggs@ubuntu:~/nw3c/therock on Instagram_ “The hierarchy of power in the _DCUniverse is ab
ople _Ruthless Shooting begins this…”.mhtml_files$ ls
_0_start_me.htm                                                   79829386_1382404908604435_
11850309_1674349799447611_206178162_a.jpg__nc_ht_s.com__nc_o.jpg  82763503_1457250774457881_
12747776_465301190322098_16890636_a.jpg__nc_ht_sco.com__nc_o.jpg  90350334_517869752483937_1
_1_info.nfo                                                       90639798_2385250921767442_
230736fb5e77.css                                                  91006776_650941012361406_5
27a959f10aa9.css                                                  92020832_600634277463000_8
2c4993169770.css                                                  92382470_880249492399923_8
31765099_218494625423607_1426631661482672128_n.jpg.com__nc_o.jpg  92565298_258098155320449_2
3be21f338c88.png                                                  92656484_2948607295186854_
41cb58da56d7.css                                                  92688026_644274202817854_5
58b93ef96160.css                                                  93122108_286959075625713_8
65027530_1019005081636040_9070353304264900608_n.jp.com__nc_o.jpg  93215725_2493094787610671_
660ee7ce042a.css                                                  93495314_669819723587229_1
66718158_463662487551057_8022966374550732800_n.jpg.com__nc_o.jpg  935344957e35.png
66774262_2537958856227088_4537248514291269632_n.jp.com__nc_o.jpg  af986a96d279.css
69280923_2487580658180164_7594294368865878016_n.jp.com__nc_o.jpg  fcd04626356d.png
69717576_425469574990885_6464646415862726656_n.jpg.com__nc_o.jpg  index.htm
```



FLAG: from Facebook

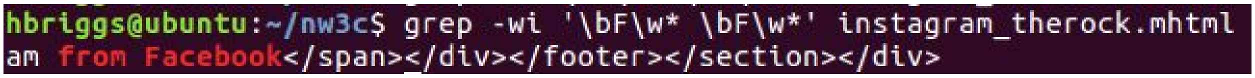
Easy, Unintended way for Added Data

The question says “two words starts with the letter F” and “in order”. Could it be possible that those words are in plain text in mhtml? Turns out the answer is “yes” because of the way the mhtml file and the website was constructed. Use Regular Expression, regex, with the following:

'\bF\w* \bF\w*'

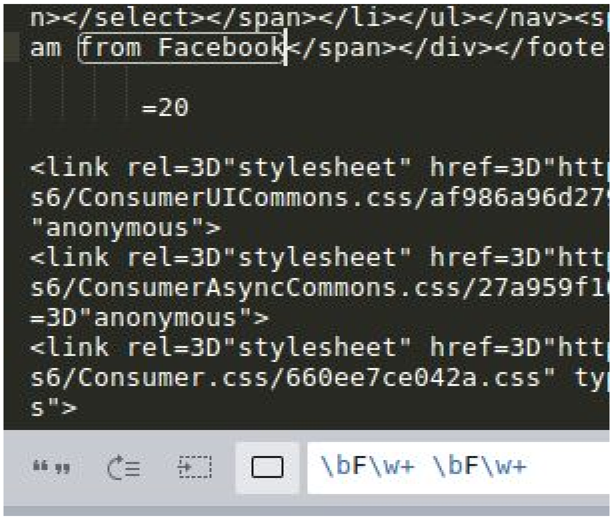
- \b to start
- F is the letter F we are focusing on
- \w+ any letter to end of a string
  - \w* was used for grep

In linux with the grep command:
grep -wi '\bF\w* \bF\w*'



```
hbriggs@ubuntu:~/nw3c$ grep -wi '\bF\w* \bF\w*' instagram_therock.mhtml
am from Facebook</span></div></footer></section></div>
```

Using Sublime3 editor, can use the built in regex editor with: \bF\w+ \bF\w+



```
n></select></span></li></ul></nav><s
am from Facebook</span></div></foote

       =20

<link rel=3D"stylesheet" href=3D"htt
s6/ConsumerUICommons.css/af986a96d27
"anonymous">
<link rel=3D"stylesheet" href=3D"htt
s6/ConsumerAsyncCommons.css/27a959f1
=3D"anonymous">
<link rel=3D"stylesheet" href=3D"htt
s6/Consumer.css/660ee7ce042a.css" ty
s">
```

\bF\w+ \bF\w+

However, this string was grouped with a whole host of other HTML tags and making it difficult to match this with the .png image without going through the HTML tags more.

Summary of Flags:

1. https://www.instagram.com/p/B8JbwzzljUb/
2. 2020-04-02 05:35:17
3. from Facebook

Summary of Questions:

- What is the URL of the saved webpage?
- When did Babacanbabajo comment on the post?
    - Put in 24hr UTC Flag Format: yyyy-mm-dd hh:mm:ss
- There is a picture inside the file that Instagram has embedded. It has two words approximately in the middle surrounded by other images/icons. They both start with the letter F. What are the two words in order starting with the one on top?

Summary of Tools:

- http://cyberchef.io
- https://sourceforge.net/projects/mht2htm/
- https://www.sublimetext.com/3
- Tar utility
- Image Editor
- Webbrowser