CSL202 | Assignment-5 | Due 5/Apr/2018 11:59 PM | 100 points

- Important instructions for coding submission are here: https://goo.gl/IMWvdF
- Grading scheme to be followed is available here: https://goo.gl/52D82g
- Assignment description may be underspecified to allow some room for exploration and creativity.
- Your submission should be packaged as a zip file named **exactly** in this format: CSL202-[your entry no.]-[assignment no.].zip.

Write a program in Python which can find duplicate hyperlinks in an HTML file, and remove the duplicates based on user's selections. An example of a hyperlink in HTML is as shown below:

```
<a href="http://example.com">Display text</a>
```

Your program should do the following:

1. Take as command-line input the HTML file in which it will find the duplicate hyperlinks.
2. Find the duplicate hyperlinks in the input file. Two hyperlinks are considered duplicate if the value of **`href`** attribute is same in both.
3. If any duplicates found then your program should print the following details:
    a. Total number of duplicates found.
    b. A serial number (e.g. 1, 2, 3 ...) for each hyperlink which is a duplicate.
    c. The hyperlink URL (i.e. the value of **`href`** attribute) and the display text (i.e. the text present between tags <a>).
    d. Line numbers in the file at which duplicates are present.
4. Your program should then allow the user to selectively remove duplicates. It should ask the user to select the hyperlinks to be retained, after which it should remove the remaining ones.
    a. Selection should be made by taking serial number that you print for the duplicates.
    b. An option should be given to the user to:
        i. Retain all the links, say, by asking the user to enter 'A'
        ii. Retain only first occurrence of a duplicate link. Say, by asking the user to enter 'F'
5. If the user has opted for removing any duplicates then the output HTML file should be written in the working directory (i.e. where the python script is executed). Output file should be named by adding a suffix ".dedup" to the input file name.

Example run:

```
$ python find_dups.py chrome_bookmarks.html
Found 5 duplicates:
   1. http://www.abc.com/tez “TEZ at its best” at line 12
   2. http://www.abc.com/tez “The new payment gateway” at line 35
   3. http://www.abc.com/tez “Pay via TEZ” at line 43
   4. http://www.andy.ss/yoyo “Playing gadget” at line 65
   5. http://www.andy.ss/yoyo “Kids toys” at line 79

Select hyperlinks that you want to keep.
       Enter A to keep all, OR
       Enter F to keep the first one in a set of duplicates, OR
       Enter the serial numbers (separated by commas) of the links to keep.
Your selection: 3, 5

Removed 3 hyperlinks. Output file written to chrome_bookmarks.html.dedup
```

HINTS:

1. You can make use of BeautifulSoup module to process the input HTML file for finding the duplicate. https://www.crummy.com/software/BeautifulSoup/bs4/doc/