Prerna Garg

2016csb1050
CSL202

1. What does this program do

a) The code uses Python's beautifulsoup package to find duplicate hyperlinks in an input html document.
b) 2 links are considered duplicate if their href attribute is same.
c) The html file is given as command line input while running the program
d) Console Output:
    - Total number of duplicates found
    - List of duplicates
    - If a link appears once, it is not printed in the output, if it appears >1 times then all its occurrences are printed.
    - Output form contains href , text , and line number of the link
    - The program then asks the user if he wishes to remove the duplicate links from the file
    - It gives 3 options i.e.
        - Do not remove (Enter A)
        - Remove all but first occurrence (Enter F)
        - Retain only the serial numbers given as input and remove the rest (Enter comma separated integers)
e) File output:
   The output is written in a new file named output.html.dedup which is saved in the same directory the code was run from.

2. A description of how this program works (i.e. its logic)

The code has been divided into 4 functions. They have been explained in the order of call as follows :

file_read - It reads the input file and finds all the links present in the file along with their line numbers, href attributes and text data. It parses the file line by line and uses beautifulsoup's find_all('a') function on every line to find the links and their corresponding line numbers.
For the following input file
<html>
<head><title>The Dormouse's story</title></head>
<body>
flipkart1
google1
flipkart2
flipkart3
</body>
</html>
The data is stored in the dictionary in the following form.
{ www.flipkart.com : (["flipkart1" , "flipkart2" , "flipkart3"] , [4 , 6 ,7] , [flipkart1 , flipkart2 , flipkart3] , [1 , 3 , 4]) ,
www.google.com : (["google1"], [5] , [google1] , [2]) }

Input: name of the input file
Output: dictionary with key as href attribute and values as list of data

a) num_dups - It calculates the number of duplicates present in the dictionary returned by file_read.
   Input: Python dictionary
   Output: integer
b) Print_dups - It will print those links to the console that have occurred more than once in the input file.

Input: Python dictionary

Output: New python dictionary with all data of those links that have occurred more than once in the input dictionary and serial numbers of all duplicate links.

c) Rem_dups - It will take user input to find what links have to be removed and perform the removal operation.

Input: input file name

Output: None

3. How to compile and run this program

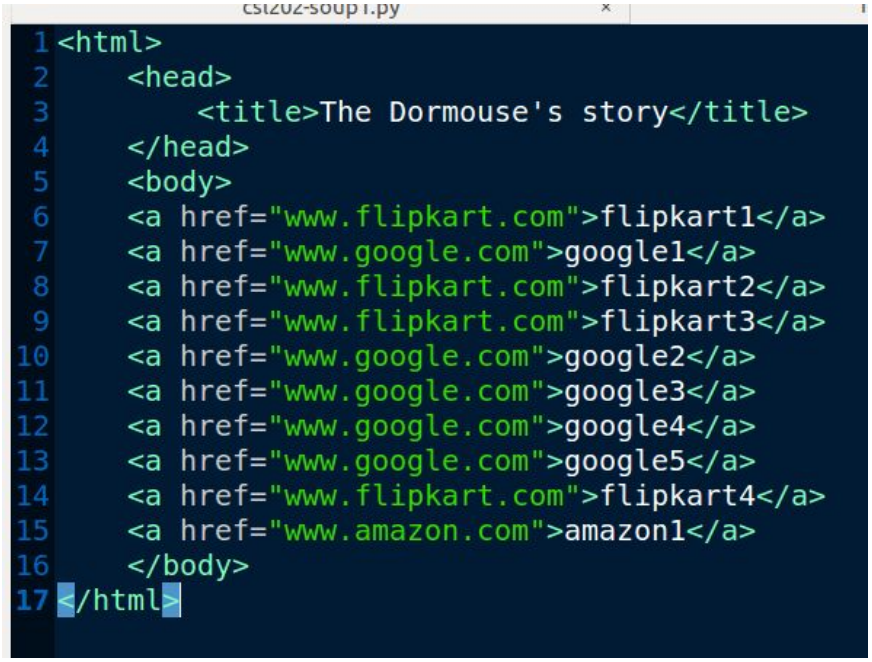
The program can be run with a single command :

python CSL202-2016csb1050-assignment5.py index.html

4. Assumptions :

1) I have assumed that no link is spread across multiple lines. However one line having multiple links has been efficiently handled.
2) If user enters invalid inputs like T,O,P then the program will terminate printing "Invalid Inputs"

4. Provide a snapshot of a sample run



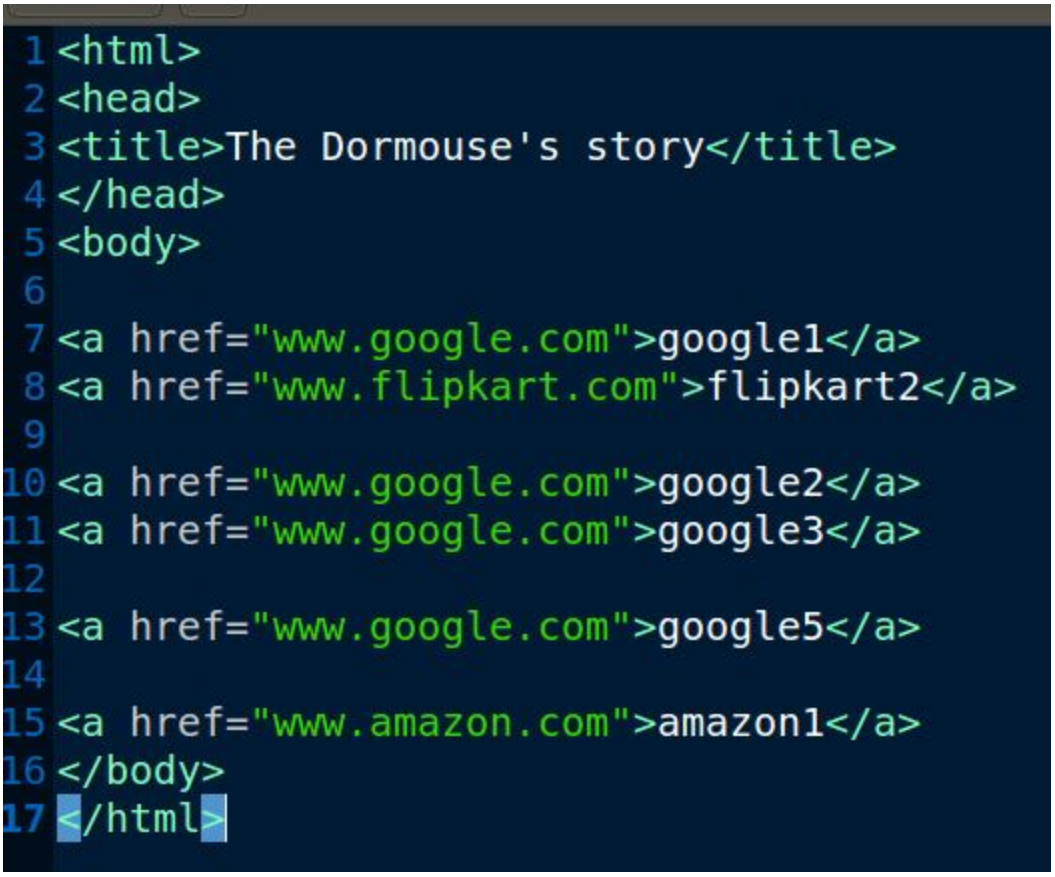
Input file

```
<html>
<head>
<title>The Dormouse's story</title>
</head>
<body>
<a href="www.flipkart.com">flipkart1</a>
<a href="www.google.com">google1</a>
<a href="www.flipkart.com">flipkart2</a>
<a href="www.flipkart.com">flipkart3</a>
<a href="www.google.com">google2</a>
<a href="www.google.com">google3</a>
<a href="www.google.com">google4</a>
<a href="www.google.com">google5</a>
<a href="www.flipkart.com">flipkart4</a>
<a href="www.amazon.com">amazon1</a>
</body>
</html>
```

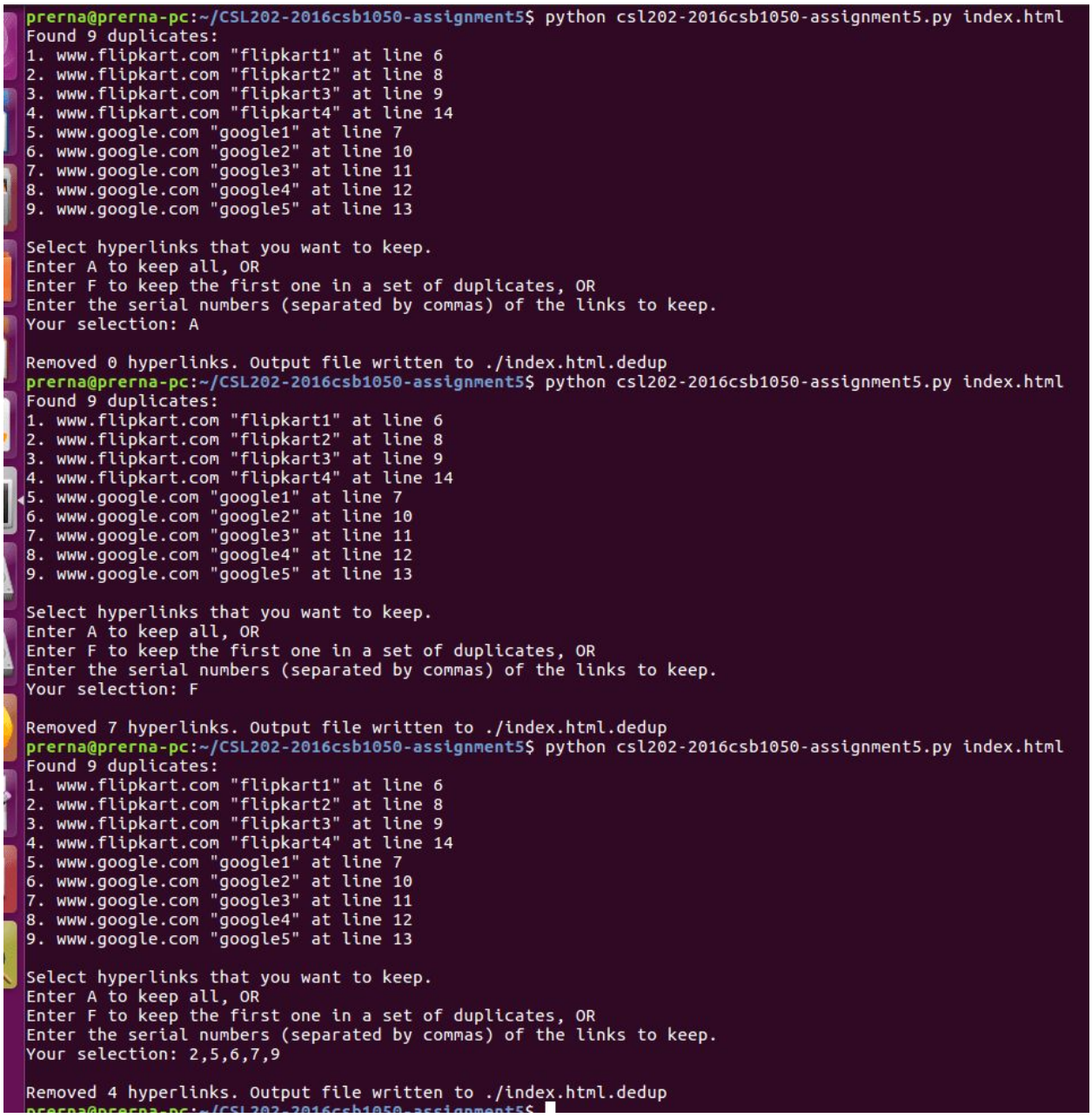
```
<html>
<head>
<title>The Dormouse's story</title>
</head>
<body>
<a href="www.flipkart.com">flipkart1</a>
<a href="www.google.com">google1</a>


<a href="www.amazon.com">amazon1</a>
</body>
</html>
```

Outputs for A and F



Output for serial numbers



Console Outputs