Assignment 2

1) The “htags1” program is designed to parse HTML-like tags from input and print them with their content, excluding any comments within. It uses a simple algorithm to achieve this. Upon encountering a '<' character, it begins reading characters into a buffer (`tag`) until it finds '>'. While reading, it ignores any characters if it's inside a comment (`inside_comment` flag). Once it reads the complete tag, it prints it along with its content. The data structure used is a character array (`tag`) to store the tag being parsed. The algorithm relies on a loop that continuously reads characters until the end of the input stream is reached.

2)

```
/**
@author Saheb Singh Arora
Student id: 3742233
*/
#include <stdio.h>

void printArr(char arr[], int size) {
    for (int i = 0; i < size; i++) {
        putchar(arr[i]);
    }

}

int main() {
    int c;
    int inside_comment = 0;
    char tag[100];

    while ((c = getchar()) != EOF) {
        if (c == '<') {
            int i = 0;
            tag[i++] = c;
            while ((c = getchar()) != EOF && c != '>') {
                if (!(inside_comment)) {
                }
                tag[i++] = c;
            }
            if (!inside_comment) {
                printArr(tag, i);
                putchar('>');
                putchar('\n');
            }
        }
    }
```

```
    return 0;
}
```

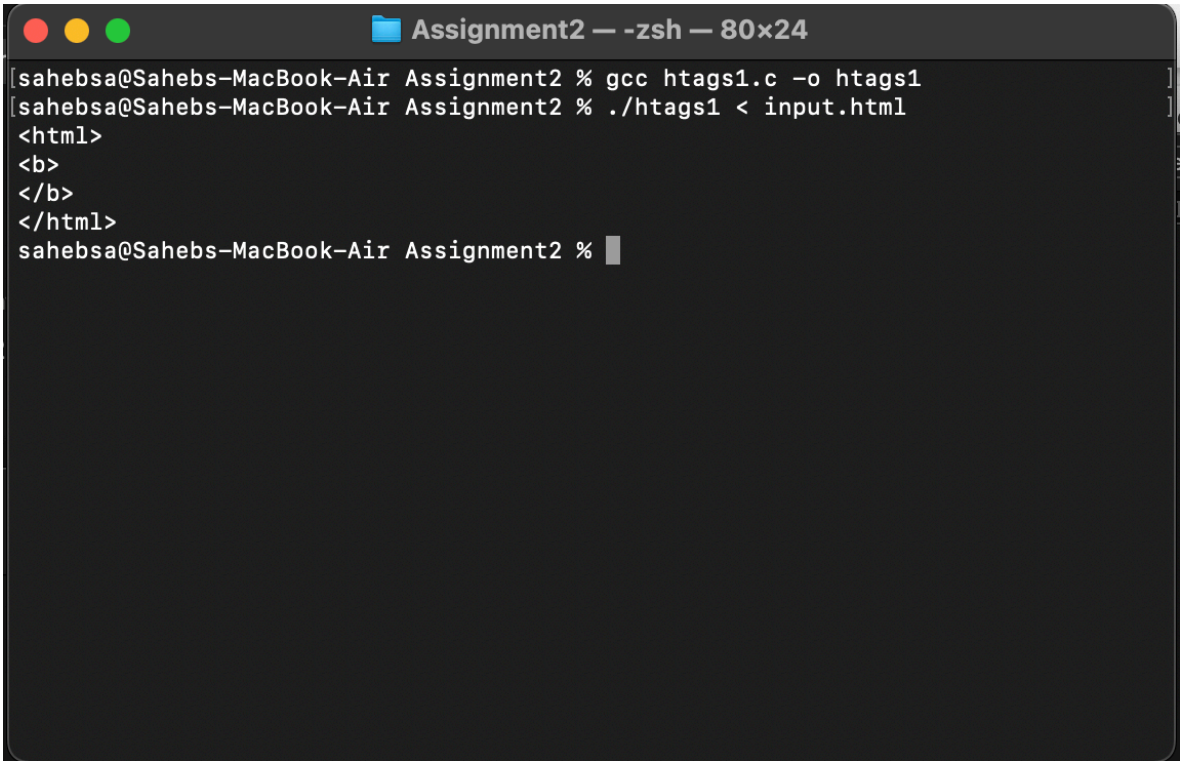
3) The “htext1” program, designed by Saheb Singh Arora, is a simple utility to strip HTML tags and comments from an input stream. It operates by iterating through each character of the input, maintaining boolean flags (`insideTag` and `insideComment`) to track whether it's currently within an HTML tag or comment. When encountering a '<' character, it distinguishes between the start of a comment or a tag; if it's a comment, it continues reading until it finds the end marker '-->'. If it's a tag, it skips reading until the closing '>'. Characters outside of tags and comments are then outputted, effectively providing a plain text representation of the input. This algorithmic approach doesn't require complex data structures, relying instead on simple flags and character iteration.

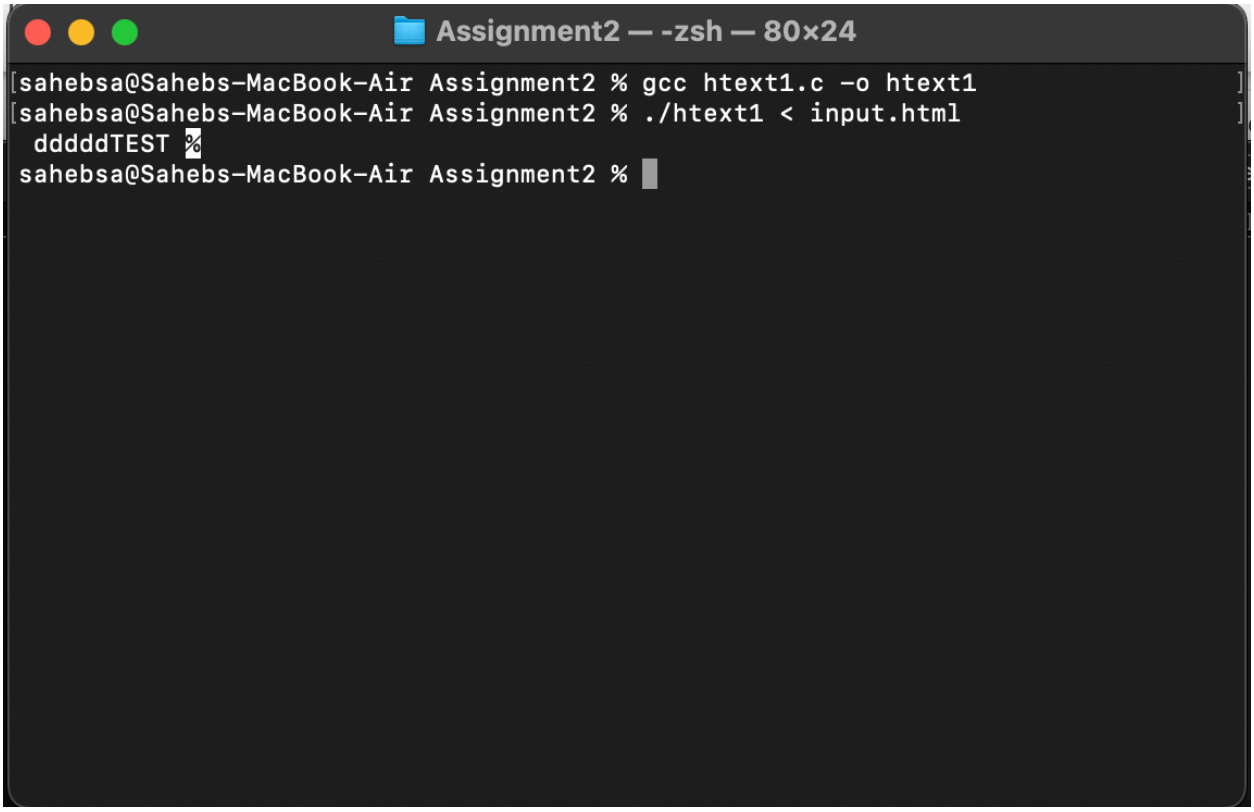
4)

```
/**
@author Saheb Singh Arora
Student id: 3742233
*/
#include <stdio.h>
#include <stdbool.h>

int main() {
    int character;
    bool insideTag = false;
    bool insideComment = false;

    for (; ;) {
        character = getchar();
        if (character == EOF)
            break;

        if (character == '<') {
            insideTag = true;
            if ((character = getchar()) == '!') {
                if ((character = getchar()) == '-' && (character
= getchar()) == '-') {
                    insideComment = true;
                    insideTag = false;
                    continue;
                } else {
```

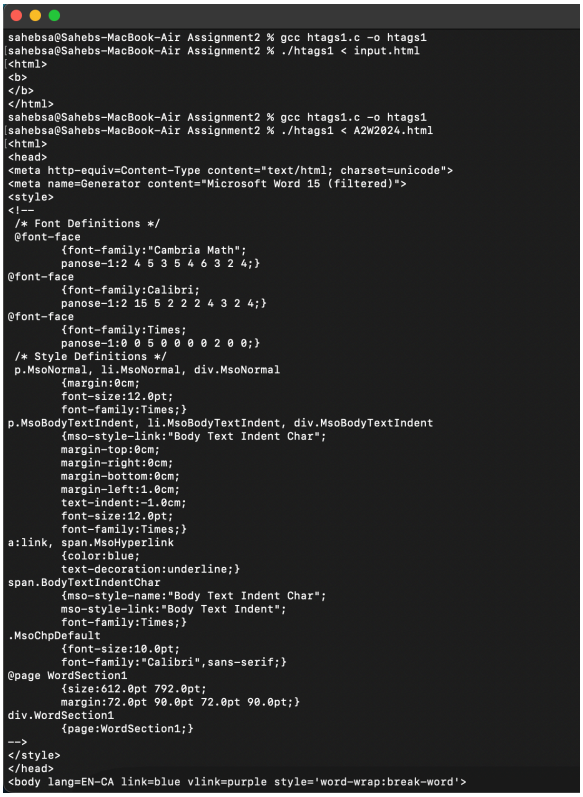
```
                    ungetc(character, stdin);
                    character = '<';
                }
            } else {
                ungetc(character, stdin);
                continue;
            }
        }
        if (character == '>') {
            if (insideComment) {
                insideComment = false;
            } else {
                insideTag = false;
            }
            continue;
        }
        if (!insideTag && !insideComment) {
            putchar(character);
        }
    }
    return 0;
}
```

5)





6)



```
sahebsa@Sahebs-MacBook-Air Assignment2 % gcc htags1.c -o htags1
sahebsa@Sahebs-MacBook-Air Assignment2 % ./htags1 < input.html
<html>
<b>
</b>
</html>
sahebsa@Sahebs-MacBook-Air Assignment2 % gcc htags1.c -o htags1
sahebsa@Sahebs-MacBook-Air Assignment2 % ./htags1 < A2W2024.html
<html>
<head>
<meta http-equiv=Content-Type content="text/html; charset=unicode">
<meta name=Generator content="Microsoft Word 15 (filtered)">
<style>
<!--
 /* Font Definitions */
 @font-face
        {font-family:"Cambria Math";
        panose-1:2 4 5 3 5 4 6 3 2 4;}
@font-face
        {font-family:Calibri;
        panose-1:2 15 5 2 2 2 4 3 2 4;}
@font-face
        {font-family:Times;
        panose-1:0 0 5 0 0 0 0 2 0 0;}
 /* Style Definitions */
 p.MsoNormal, li.MsoNormal, div.MsoNormal
        {margin:0cm;
        font-size:12.0pt;
        font-family:Times;}
p.MsoBodyTextIndent, li.MsoBodyTextIndent, div.MsoBodyTextIndent
        {mso-style-link:"Body Text Indent Char";
        margin-top:0cm;
        margin-right:0cm;
        margin-bottom:0cm;
        margin-left:1.0cm;
        text-indent:-1.0cm;
        font-size:12.0pt;
        font-family:Times;}
a:link, span.MsoHyperlink
        {color:blue;
        text-decoration:underline;}
span.BodyTextIndentChar
        {mso-style-name:"Body Text Indent Char";
        mso-style-link:"Body Text Indent";
        font-family:Times;}
.MsoChpDefault
        {font-size:10.0pt;
        font-family:"Calibri",sans-serif;}
@page WordSection1
        {size:612.0pt 792.0pt;
        margin:72.0pt 90.0pt 72.0pt 90.0pt;}
div.WordSection1
        {page:WordSection1;}
-->
</style>
</head>
<body lang=EN-CA link=blue vlink=purple style='word-wrap:break-word'>
```

```
</style>
</head>
<body lang=EN-CA link=blue vlink=purple style='word-wrap:break-word'>
<div class=WordSection1>
<p class=MsoNormal>
<b>
<span lang=EN-US style='font-size:14.0pt;font-family:
"Times New Roman",serif'>
</span>
</b>
</p>
<p class=MsoNormal>
<b>
<span lang=EN-US style='font-size:14.0pt;font-family:
"Times New Roman",serif'>
</span>
</b>
</p>
<p class=MsoNormal>
<b>
<span lang=EN-US style='font-size:14.0pt;font-family:
"Times New Roman",serif'>
</span>
</b>
</p>
<p class=MsoNormal style='text-align:justify'>
<span lang=EN-US
style='font-family:"Times New Roman",serif'>
</span>
</p>
<p class=MsoNormal>
<u>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
<b>
</b>
</span>
</u>
</p>
<p class=MsoNormal>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
</p>
<p class=MsoNormal>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
<a href="https://www.educba.com/types-of-tags-in-html/">
<span
style='text-decoration:none'>
</span>
</a>
</span>
</p>
<p class=MsoNormal>
<span style='font-family:"Times New Roman",serif'>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:0cm;margin-bottom:.0001pt;text-indent:0cm;line-height:12.0pt'>
```

```
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:0cm;margin-bottom:.0001pt;text-indent:0cm;line-height:12.0pt'>
<i>
<span
style='font-family:"Times New Roman",serif'>
</span>
</i>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:7.1pt;margin-bottom:.0001pt;text-indent:0cm;line-height:12.0pt'>
<span
lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
<span
lang=EN-US style='font-family:"Courier New"'>
</span>
<span lang=EN-US
style='font-family:"Times New Roman",serif'>
<b>
</b>
</span>
<span lang=EN-US style='font-family:"Courier New"'>
</span>
<span
lang=EN-US style='font-family:"Times New Roman",serif'>
<a
href="https://linuxcommand.org/lc3_lts0070.php">
</a>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:7.1pt;margin-bottom:.0001pt;text-indent:18.0pt;line-height:
12.0pt'>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:25.1pt;margin-bottom:.0001pt;text-indent:18.0pt;line-height:
12.0pt'>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
<br>
<br>
<br>
<br>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:25.1pt;margin-bottom:.0001pt;text-indent:18.0pt;line-height:
12.0pt'>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
<br>
<br>
<br>
<br>
<br>
<br>
```

```
12.0pt'>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
<br>
<br>
<br>
<br>
<br>
<br>
<br>
<br>
<br>
<br>
<br>
<br>
<br>
<br>
<br>
<br>
<br>
<br>
<br>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:7.1pt;margin-bottom:.0001pt;text-indent:0cm;line-height:12.0pt'>
<span
lang=EN-US style='font-family:"Times New Roman",serif'>
<br>
<br>
<br>
</span>
<b>
<span lang=EN-US
style='font-family:"Courier New"'>
</span>
</b>
<span lang=EN-US
style='font-family:"Times New Roman",serif'>
</span>
<b>
<span lang=EN-US style='font-family:"Courier New"'>
</span>
</b>
<span
lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
<b>
<span lang=EN-US style='font-family:"Courier New"'>
</span>
</b>
<span
lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
<span lang=EN-US style='font-family:"Courier New"'>
</span>
<span
```

```
<span lang=EN-US style='font-family:"Courier New"'>
</span>
</b>
<span
lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
<b>
<span lang=EN-US style='font-family:"Courier New"'>
</span>
</b>
<span
lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
<span lang=EN-US style='font-family:"Courier New"'>
</span>
<span
lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:43.1pt;margin-bottom:.0001pt;text-indent:-18.0pt;line-height:
12.0pt'>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
<span
lang=EN-US style='font-size:7.0pt;font-family:"Times New Roman",serif'>
</span>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:43.1pt;margin-bottom:.0001pt;text-indent:-18.0pt;line-height:
12.0pt'>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
<span
lang=EN-US style='font-size:7.0pt;font-family:"Times New Roman",serif'>
</span>
<span style='font-family:"Times New Roman",serif'>
<u>
</u>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:43.1pt;margin-bottom:.0001pt;text-indent:-18.0pt;line-height:
12.0pt'>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
<span
lang=EN-US style='font-size:7.0pt;font-family:"Times New Roman",serif'>
</span>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:43.1pt;margin-bottom:.0001pt;text-indent:-18.0pt;line-height:
12.0pt'>
```

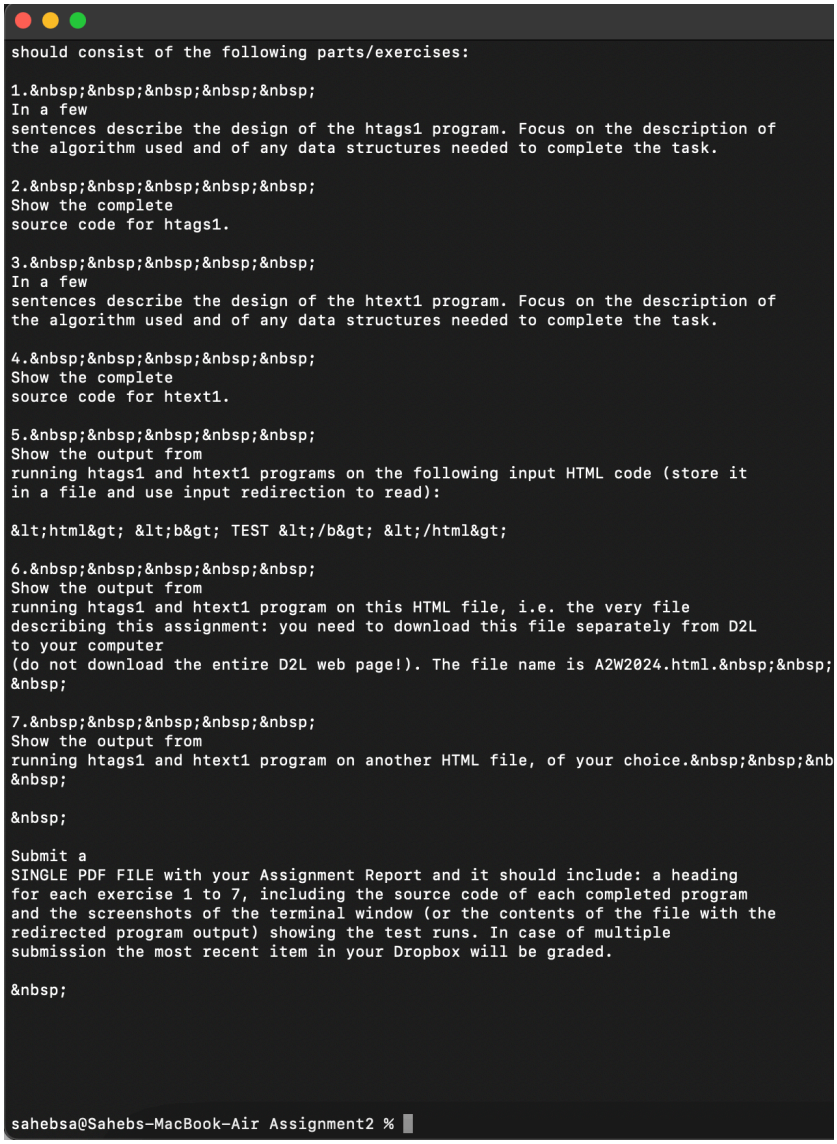
```
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:43.1pt;margin-bottom:.0001pt;text-indent:-18.0pt;line-height:
12.0pt'>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
<span
lang=EN-US style='font-size:7.0pt;font-family:"Times New Roman",serif'>
</span>
<span style='font-family:"Times New Roman",serif'>
<u>
</u>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
12.0pt;margin-left:43.1pt;text-indent:-18.0pt;line-height:12.0pt'>
<span
lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
<span
lang=EN-US style='font-size:7.0pt;font-family:"Times New Roman",serif'>
</span>
<span style='font-family:"Times New Roman",serif'>
<br>
<br>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:43.1pt;margin-bottom:.0001pt;text-indent:-18.0pt;line-height:
12.0pt'>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
<span
lang=EN-US style='font-size:7.0pt;font-family:"Times New Roman",serif'>
</span>
<span style='font-family:"Times New Roman",serif'>
</span>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:43.1pt;margin-bottom:.0001pt;text-indent:-18.0pt;line-height:
12.0pt'>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
<span
lang=EN-US style='font-size:7.0pt;font-family:"Times New Roman",serif'>
</span>
<span style='font-family:"Times New Roman",serif'>
</span>
<span
lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;text-indent:0cm;line-height:
```

```
<span
lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
<span
lang=EN-US style='font-size:7.0pt;font-family:"Times New Roman",serif'>
</span>
<span style='font-family:"Times New Roman",serif'>
<br>
<br>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:43.1pt;margin-bottom:.0001pt;text-indent:-18.0pt;line-height:
12.0pt'>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
<span
lang=EN-US style='font-size:7.0pt;font-family:"Times New Roman",serif'>
</span>
<span style='font-family:"Times New Roman",serif'>
</span>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;margin-right:0cm;margin-bottom:
0cm;margin-left:43.1pt;margin-bottom:.0001pt;text-indent:-18.0pt;line-height:
12.0pt'>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
<span
lang=EN-US style='font-size:7.0pt;font-family:"Times New Roman",serif'>
</span>
<span style='font-family:"Times New Roman",serif'>
</span>
<span
lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;text-indent:0cm;line-height:
12.0pt'>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
</p>
<p class=MsoBodyTextIndent style='margin-top:6.0pt;text-indent:0cm;line-height:
12.0pt'>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
</p>
<p class=MsoNormal>
<span lang=EN-US style='font-family:"Times New Roman",serif'>
</span>
</p>
</div>
</body>
</html>
zsh: abort      ./htags1 < A2W2024.html
sahebsa@Sahebs-MacBook-Air Assignment2 %
```

Using htext1

```
[sahebsa@Sahebs-MacBook-Air Assignment2 % gcc htext1.c -o htext1
[sahebsa@Sahebs-MacBook-Air Assignment2 % ./htext1 < A2W2024.html
??









CS 2263

Assignment 2

Due: Feb. 21, 2024 (before midnight)

---------------------------------------------------------

Assignments
are to be completed individually (not as a group work).

 

Sample
reference: https://www.educba.com/types-of-tags-in-html/


 

Processing strings

Develop and test two
utility programs called  htags1  and  htext1  to process an HTML file. The htags1 prog
HTML tags, except the HTML comment tags &lt;!-- --&gt;, one per line of
output. The htext1 program should print
all plain text extracted from the source HTML file, i.e. just the text, no
tags. The utility programs should read form standard input and may use the input redir
a file. (This means that your program code should read from standard input, and
NOT perform any file operations, like fopen() etc. You may capture the output
from your program using output redirection, or you may just output to the
terminal window and print the screen, your choice.)

Assume an
HTML tag is any text enclosed between &lt; and &gt;, including these delimiter
characters. The end tags, like &lt;/p&gt;, are valid tags. For example, in a
sample input HTML file containing:


&lt;body lang=EN-CA link=blue vlink=&quot;#954F72&quot;&gt;
&lt;div class=WordSection1&gt;
```

```
&lt;body lang=EN-CA link=blue vlink=&quot;#954F72&quot;&gt;
&lt;div class=WordSection1&gt;
&lt;p class=MsoNormal&gt;&lt;b&gt;&lt;span lang=EN-US
style='font-size:14.0pt;font-family: &quot;Times New Roman&quot;,serif'&gt;CS
2263&lt;/span&gt;&lt;/b&gt;&lt;/p&gt;
&lt;p class=MsoNormal&gt;&lt;b&gt;&lt;span lang=EN-US
style='font-size:14.0pt;font-family: &quot;Times New
Roman&quot;,serif'&gt;Assignment 2&lt;/span&gt;&lt;/b&gt;&lt;/p&gt;


we find the following tags:

&lt;body lang=EN-CA link=blue vlink=&quot;#954F72&quot;&gt;
&lt;div class=WordSection1&gt;
&lt;p class=MsoNormal&gt;
&lt;b&gt;
&lt;span lang=EN-US style='font-size:14.0pt;font-family: &quot;Times New
Roman&quot;,serif'&gt;
&lt;/span&gt;
&lt;/b&gt;
&lt;/p&gt;
&lt;p class=MsoNormal&gt;
&lt;b&gt;
&lt;span lang=EN-US style='font-size:14.0pt;font-family: &quot;Times New
Roman&quot;,serif'&gt;
&lt;/span&gt;
&lt;/b&gt;
&lt;/p&gt;

and the plain text in this sample input is:

CS 2263 Assignment 2


When outputting the plain text you may keep new line characters present in the
input, or you may replace them with space characters.

The data to your program should be read from standard input (use the input
redirection to read from the HTML file) using the getchar() function. The output shoul
done using the putchar() function and/or using
the printArr() function from Assignment
1. You must use the printArr() function at least once in
each utility program developed for this assignment. The Assignment Report
should consist of the following parts/exercises:

1.     
In a few
sentences describe the design of the htags1 program. Focus on the description of
the algorithm used and of any data structures needed to complete the task.

2.     
Show the complete
source code for htags1.

3.     
In a few
sentences describe the design of the htext1 program. Focus on the description of
```


should consist of the following parts/exercises:
1.
In a few
sentences describe the design of the htags1 program. Focus on the description of
the algorithm used and of any data structures needed to complete the task.
2.
Show the complete
source code for htags1.
3.
In a few
sentences describe the design of the htext1 program. Focus on the description of
the algorithm used and of any data structures needed to complete the task.
4.
Show the complete
source code for htext1.
5.
Show the output from
running htags1 and htext1 programs on the following input HTML code (store it
in a file and use input redirection to read):
<html> TEST </html>
6.
Show the output from
running htags1 and htext1 program on this HTML file, i.e. the very file
describing this assignment: you need to download this file separately from D2L
to your computer
(do not download the entire D2L web page!). The file name is A2W2024.html.

7.
Show the output from
running htags1 and htext1 program on another HTML file, of your choice. &nb

Submit a
SINGLE PDF FILE with your Assignment Report and it should include: a heading
for each exercise 1 to 7, including the source code of each completed program
and the screenshots of the terminal window (or the contents of the file with the
redirected program output) showing the test runs. In case of multiple
submission the most recent item in your Dropbox will be graded.

sahebsa@Sahebs-MacBook-Air Assignment2 %