CSCE 221 Assignment 4 Cover Page

First Name Leilani Last Name Horlander-Cruz UIN 523008771

User Name leilanihc112 E-mail address leilanihc112@tamu.edu

Please list all sources in the table below including web pages which you used to solve or implement the current homework. If you fail to cite sources you can get a lower number of points or even zero, read more on Aggie Honor System Office website: http://aggiehonor.tamu.edu/

Type of sources				
People	Ian Dickerson			
Web pages (provide URL)	cplusplus.com			
Printed material				
Other Sources				

I certify that I have listed all the sources that I used to develop the solutions/codes to the submitted work.

On my honor as an Aggie, I have neither given nor received any unauthorized help on this academic work.

Your Name	Leilani		Horlander-Cruz	Date	3-30-17

\begin_inset Separator latexpar\end_inset

CSCE 221 — Programming Assignment 4, Spring 2017 (100 points)

Due: March 30th at 11:59 pm

File Parsing and Regex – Assignment Description (100 points)

A very common task in Computer Science is the reading in and parsing of text. One of the most powerful tools at a programmer's disposal to aid in this task is regex (which stands for REGular EXpressions).

1. Read about text manipulation and regex: Programming Chap. 2&3 (PPT)
2. Reference for C++ regex grammar: regex reference
3. (15 points) Compile and run the following code, then answer the questions:
    1. What is stored in "matches"?
        1. A string that matches the parameters of two digits together. In this case, "98" is stored in "matches".
    2. What does "\d" mean?
        1. Digit

```
#include <iostream>
#include <string>
#include <regex>
using namespace std;
int main() {
   regex pattern{R"(\d\d)"};
   string to_search = "I would like the number 98"
              " to be found and printed, thanks.";
   smatch matches;
   regex_search(to_search, matches, pattern);
   for (auto match : matches) {
      cout << match << endl;
   }
   return 0;
}
```

    Note that in C/C++ two subsequent strings as above are combined into one string by the compiler.
    3. Modify the regex pattern to retrieve a two-digit number and the word thanks in the string. Test your pattern for correctness.
4. (25 points) Compile and run the following code, then answer the questions:
    1. What does "\s\S" mean?
        1. A character set that matches any character including a newline
    2. What is stored in matches[0]?
        1. <title> This is a title </title>
    3. Why is matches[1] different?
        1. It stores the ([\s\S]+) in matches[1] as a unique match from matches[0]

```
#include <iostream>
#include <string>
#include <regex>
using namespace std;
int main() {
   regex pattern{R"(<title>([\s\S]+)</title>)"};
   string to_search = "<html><head>Wow such a header
<title>This is a title</title>"
                      "So top</head>Much body</html>";
```

```
    smatch matches;
    regex_search(to_search, matches, pattern);
    cout << matches[0] << endl;
    cout << matches[1] << endl;
    return 0;
}
```

4. Modify the regex pattern to retrieve only the items inside of the header tag but not inside of the title tag. Test your pattern for correctness.

5. (40 points) Download the following text file: stroustrup.txt
   Write a program using regex that will go through the text file and print out the file name of every hyperlinked powerpoint file. (Hint1: an HTML hyperlink uses the format .... Hint2. The powerpoint file extension is .ppt)

6. Description of data structures and algorithms used by your program.
   Regular expression data structure is a text string for describing a search pattern.
7. Description of input and output data. List all restrictions and assumptions that you have imposed on your input data and program.
   The .txt file provided to parse was the only input file used. No other input methods were used. The output of the program for the first problem was "98 thanks", for the second was "Wow such a header So top", and for the last was
   "1-2_programming.ppt
   3_types.ppt
   4_computation.ppt
   5_errors.ppt
   6_writing.ppt
   7_completing.ppt
   8_functions.ppt
   9_classes.ppt
   10_iostreams.ppt
   11_custom_io.ppt
   12_display.ppt
   13_graph_classes.ppt
   14_class_design.ppt
   15_graphing.ppt
   16_GUI.ppt
   17_free_store.ppt
   18_arrays.ppt
   19_vector.ppt
   20_containers.ppt
   21_algorithms.ppt
   22_ideals.ppt
   23_text.ppt
   24_numerics.ppt
   25_embedded.ppt
   26_testing.ppt
   27_C.ppt
   Review_for_Final.ppt"

8. Write your regex patterns used for parsing the strings in the programs above. Explain their syntax.
   For problem 3:
   regex pattern{R”(\d\d)”};
   regex pattern2{”thanks”};
   - pattern has “\d\d” which means a substring pf two digits back to back in the full string
   - pattern2 has “thanks” which just finds the word “thanks” in the string
     For problem 4:regex pattern{R"(<head>(.*)<title>)"};regex pattern2{R"(</title>(.*)</head>)"};
   - pattern has “<head>(.*)<title>” which means any characters including and between <head> and <title>
   - pattern2 has “</title>(.*)</head>” which means any characters including and between </title> and </head>
     For problem 5:regex pattern{ R"((.*))" };
   - pattern has “((.*))” which means an HTML hyperlink with any characters followed by “.ppt”
9. What is the purpose of the functions regex_search() and regex_match().
   Regex_search returns whether a sub-sequence in the target sequence (the subject) matches the regular expression (the pattern). Regex_match returns the target sequence matches the regular expression.
10. Which C++ features or standard library classes have you used in your program?
    #include <iostream>
    #include <string>
    #include <regex>
    #include <fstream>
11. Write your conclusion.
    This programming assignment helped better my understanding of regular expressions. The tasks were clear and concise and left little room for error. The outputs all came out as expected and few errors occurred.

What to submit to CSNet?

- Your C++ source code with the header block including: your name, user name, section number and e-mail address
- (20 points) A report which should consists of the following parts:
  - The cover page.
  - Assignment number and its description.
  - Description of data structures and algorithms used by your program.
  - Description of input and output data. List all restrictions and assumptions that you have imposed on your input data and program.
  - Write your regex patterns used for parsing the strings in the programs above. Explain their syntax.
  - What is the purpose of the functions regex_search() and regex_match().
  - Which C++ features or standard library classes have you used in your program?
  - Write your conclusion.