AIM 5001 Project 1 (M5) (100 Points)

Text Processing

****You may work in small groups of no more than three (3) people for this Project. ****

Part 1: Regular Expressions (30 Points)

Text data is often in need of "cleaning" and preparation before it can be effectively used for analysis purposes. Consider the following poorly formatted text string containing names and phone numbers of some residents of the town of Springfield:

```
"555-1239Dr. Anthony Fauci(636) 555-0113Hollingdorp, Donnatella555-6542Fitzgerald, F. Sco
tt555 8904Rev. Martin Luther King636-555-3226Snodgrass, Theodore5553642Carlamina Scarfoni
"
```

Use your Python regular expression ("regex") skills to complete the following tasks:

1. (**6 Points**) Extract the names of each individual from the unformatted text string shown above and store them in a vector of some sort. When complete, your vector should contain the following entries:

```
"Dr. Anthony Fauci"   "Hollingdorp, Donnatella"  "Fitzgerald, F. Scott"
"Rev. Martin Luther King"   "Snodgrass, Theodore"   "Carlamina Scarfoni"
```

2. Using your new vector containing only the names of the six individuals, complete the following tasks:

a. (**4 Points**) Use your regex skills to rearrange the vector so that all elements conform to the standard "firstname lastname", preserving any titles (e.g., "Rev.", "Dr.", etc) or middle/second names.

b. (**4 Points**) Using your regex skills, construct a logical vector indicating whether a character has a title (i.e., Rev. and Dr.).

c. (**4 Points**) Using your regex skills, construct a logical vector indicating whether a character has a middle/second name.

3. (**6 Points**) Consider the HTML string <title>+++BREAKING NEWS+++<title>. We would like to extract the first HTML tag (i.e., "<title>"). To do so we write the regular expression "**<.+>**". Explain why this fails and correct the expression.

4. (**6 Points**) Consider the string "(5-3)^2=5^2-2*5*3+3^2" conforms to the binomial theorem. We would like to extract the equation in its entirety from the string. To do so we write the regular expression "**[^0-9=+*()]+"**. Explain why this fails and correct the expression.

Part 2: Analyzing Chess Tournament Results (70 Points)

For Part 2 of Project 1, you're given a text file ("Project1.txt") with chess tournament results where the information has some structure. Your job is to create a Jupyter Notebook that generates a .CSV file with the following information for each of the chess players:

Player's Name, Player's State, Total Number of Points, Player's Pre-Rating, and Average Pre Tournament Chess Rating of Opponents

(Project Continues on Next Page)

For the first player shown in the file excerpt below, that information would be:

Gary Hua, ON, 6.0, 1794, 1605

His "Average Pre Tournament Chess Rating of Opponents" score of 1605 was calculated by using the pre-tournament opponents' ratings of 1436, 1563, 1600, 1610, 1649, 1663, 1716, and dividing by the total number of games played.

For each player we are provided with the total points they've won during the tournament and details on the results of their seven rounds of play. For each round we are given the unique ID of their opponent (an integer value) and an indicator of whether they won ('W'), lost ('L'), achieved a draw ('D'), had a bye for that round ('B'), or were unable to compete ('U').

If you have questions about the meaning of the remainder of the data or the results, please post them in the weekly Discussion Forum. Data science, like chess, is a game of back and forth…

The chess rating system (invented by a Minnesota statistician named Arpad Elo) has been used in many other contexts, including assessing the relative strength of employment candidates by human resource departments.

Be sure to include some commentary **in formatted Markdown cells** explaining your approach to solving each of the individual problems. Save all of your work for this project within **a single Jupyter Notebook** upload / submit it within the provided Project 1 Canvas submission portal. Be sure to save your Notebook using the following nomenclature : **first initial_last name_project1**" (e.g., J_Smith_project1). ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member should submit their own copy of the team's work within Canvas.***

Excerpt from text file:

```
-----------------------------------------------------------------------------------------
 Pair | Player Name                     |Total|Round|Round|Round|Round|Round|Round|Round|
 Num  | USCF ID / Rtg (Pre->Post)       | Pts |  1  |  2  |  3  |  4  |  5  |  6  |  7  |
-----------------------------------------------------------------------------------------
   1  | GARY HUA                        |6.0  |W  39|W  21|W  18|W  14|W   7|D  12|D   4|
  ON  | 15445895 / R: 1794   ->1817     |N:2  |W    |B    |W    |B    |W    |B    |W    |
-----------------------------------------------------------------------------------------
   2  | DAKSHESH DARURI                 |6.0  |W  63|W  58|L   4|W  17|W  16|W  20|W   7|
  MI  | 14598900 / R: 1553   ->1663     |N:2  |B    |W    |B    |W    |B    |W    |B    |
-----------------------------------------------------------------------------------------
   3  | ADITYA BAJAJ                    |6.0  |L   8|W  61|W  25|W  21|W  11|W  13|W  12|
  MI  | 14959604 / R: 1384   ->1640     |N:2  |W    |B    |W    |B    |W    |B    |W    |
-----------------------------------------------------------------------------------------
   4  | PATRICK H SCHILLING             |5.5  |W  23|D  28|W   2|W  26|D   5|W  19|D   1|
  MI  | 12616049 / R: 1716   ->1744     |N:2  |W    |B    |W    |B    |W    |B    |B    |
-----------------------------------------------------------------------------------------
   5  | HANSHI ZUO                      |5.5  |W  45|W  37|D  12|D  13|D   4|W  14|W  17|
  MI  | 14601533 / R: 1655   ->1690     |N:2  |B    |W    |B    |W    |B    |W    |B    |
-----------------------------------------------------------------------------------------
```

***Part 2 Grading Rubric**:* you'll receive **up to 50 points** if you successfully write the player name and total points into a pandas DataFrame, then into a .CSV file. You'll receive **up to 60 points** if you also successfully process the information from the second line for each player: state and

pre-tournament rating. **To get the full 70 points for Part 2**, you will also need to successfully calculate and process the average pre-tournament rating for each player's opponents.