DAV 5400 Fall 2019 Week 8 Assignment (30 Points)

Regular Expressions

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:

```
"555-1239Moe Szyslak(636) 555-0113Burns, C. Montgomery555 -6542Rev. Timothy Lovejoy555
8904Ned Flanders636-555-3226Simpson, Homer5553642Dr. Julius Hibbert"
```

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

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

```
"Moe Szyslak"   "Burns, C. Montgomery"  "Rev. Timothy Lovejoy"
"Ned Flanders"   "Simpson,Homer"        "Dr. Julius Hibbert"
```

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

> **a.** 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.** Construct a logical vector indicating whether a character has a title (i.e., Rev. and Dr.).
>
> **c.** Construct a logical vector indicating whether a character has a middle/second name.

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

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 assignment within **a single Jupyter Notebook** and upload it to your online DAV5400 GitHub directory. Be sure to save your Notebook using the nomenclature we've been using, i.e., **first initial_last name_W8_assn**" (e.g., J_Smith_W8_assn_).

As a reminder, this assignment is due no later than 11.59pm on Sunday Nov 3.