AIM 5001 M5 Assignment (100 Points)

Regular Expressions + String Processing

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 information on the first five presidents of the United States:

```
"*B/D1732-1799George Washington---PARTY---Unaffiliated—-SERVED:1789 – 1797, VP = John Ada
ms;#?!*****B/D1735-1826John Adams---PARTY---Federalist—-SERVED:1797 – 1801,VP:  Thomas Je
fferson;*****B/D1743-1826Thomas Jefferson ---PARTY---Democratic-Republican —-SERVED:1801
– 1809,  VP = Aaron Burr, George Clinton;##*****B/D1751-1836James Madison ---PARTY---Demo
cratic-Republican —-SERVED:1809 – 1817,  VP = George Clinton, Elbridge Gerry;?!C****B/D17
58-1831James Monroe ---PARTY---Democratic-Republican —-SERVED:1817 – 1825,  VP = Daniel D
Tompkins;"
```

Within the text string we are provided with the following information for each of the five presidents:

- **Year Born / Year Deceased**: Prefaced with '**B/D'**
- **Name**: Follows Year Born / Year Deceased data
- **Name of Political Party**: Prefaced by the word "**PARTY**"
- **Start/End of service as President**: Prefaced by the word "**SERVED**"
- **Name(s) of Vice Presidents who served during the President's term(s) in office**: Prefaced with '**VP**'. Note that more than one Vice President may have served any given President.

Use **Python regular expressions** ("regex") along with your knowledge of Python list and dictionary objects to complete the following tasks:

1. (**10 Points**) Using regular expressions, extract the **Year Born / Year Deceased** for each President from the unformatted text string shown above and store them in two separate Python list objects, i.e., one list containing the **Year Born** values and one list containing the **Year Deceased** values.

2. (**10 Points**) Using regular expressions, extract the **Name** of each President from the unformatted text string shown above and store the extracted names in a Python list object.

When complete, your list should contain the following entries:

```
"George Washington"   "John Adams"  "Thomas Jefferson"
"James Madison"   "James Monroe"
```

3. (**10 Points**) Using regular expressions, extract the **Name of Political Party** for each President from the unformatted text string shown above and store the extracted political party names in a Python list object.

4. (**10 Points**) Using regular expressions, extract the **Name(s) of Vice Presidents** for each President from the unformatted text string shown above and store the extracted names in a Python dictionary object wherein the **key:value** pairs are created using the name of each of the first five Presidents of the United States as the dictionary's key values and the names of their associated vice presidents being instantiated as data values for each President. Note that only one **key:value** pair should appear within the resulting dictionary for each President (i.e., one entry for each President).

5. Using your newly created list and dictionary objects, complete the following tasks:

> **a.** (**10 Points**) Use your regex and string processing skills to rearrange the content of the list of names of Presidents so that all elements conform to the standard "last name, first name"; then, arrange the list in alphabetical order on the basis of the first letter of the last name of each president.

b. (**10 Points**) Use your Python skills to create a new dictionary object containing **the total duration of each President's lifespan**. The resulting dictionary object should use the name of each president as **key** values while their lifespan is used to populate the associated data values for each **key:value** pair within the dictionary. Then, using your new dictionary object, calculate the AVERAGE lifespan of the first five Presidents of the United States.

c. (**10 Points**) Using your regex skills and the dictionary object created in Question 4 (above), construct a new dictionary object indicating whether each Vice President who served during the terms of the first five Presidents of the United States has either a '**G**' or a '**J**' anywhere within their first name. The resulting dictionary should be comprised of one entry for each Vice President, wherein the key value is the Vice President's name and the associated data value contains either the Python keyword '**TRUE**' or the Python keyword '**FALSE**'.

d. (**10 Points**) Using your regex skills and the dictionary object created in **Question 4** (above), construct a new dictionary object indicating whether each Vice President who served during the terms of the first five Presidents of the United States has a middle/second name or middle initial. The resulting dictionary should be comprised of one entry for each Vice President, wherein the key value is the Vice President's name and the associated data value contains either the Python keyword '**TRUE**' or the Python keyword '**FALSE**'.

6. (**10 Points**) Consider the character string '**FIdD1E7h=**'. We would like to match this string using the regular expression "**[a-zA-Z]*[^,]=**", but the regular expression fails to match the text string. Explain why the regular expression fails and correct it.

7. (**10 Points**) Consider the character string "**The spy was carefully disguised**". We would like to extract only the adverb '**carefully**' from the string. To do so we write the regular expression "**^D\s+ly()+"**. 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 / submit it within the provided Module 5 Canvas submission portal. Be sure to save your Notebook using the following nomenclature: **first initial_last name_M5_Assn**" (e.g., J_Smith_project1).