Class Project - II - Data Wrangling from multiple data sources

Due Nov 20, 2017 by 8am **Points** 20 **Submitting** a file upload
File Types png and ipynb **Available** Nov 3, 2017 at 8am - Dec 4, 2017 at 8am about 1 month

This assignment was locked Dec 4, 2017 at 8am.

You are provided with patient data as a XML file - dataset.xml. In the XML file, each patient data is within <Record></Record>.

The patient data contains MRN, and CUI of medications.

Create a CSV file with patient mrn, medication cui, medication name, medication class, medication mechanism of action, medication diagnosis.

You are also provided with UMLS data data for obtaining medication related information.

Follow the instructions of Homework 10 to help you read the UMLS data.

All data is lcoated in /opt/class/umls folder of class VM.

Read UMLS RRF documentation: **https://www.ncbi.nlm.nih.gov/books/NBK9685/** **(https://www.ncbi.nlm.nih.gov/books/NBK9685/)**

Read and familiarize with abbreviations used in the files: **https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/abbreviations.html** **(https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/abbreviations.html)**

Review the video on working with UMLS data, and homework specific instructions: **https://uthvideo.uth.tmc.edu/Panopto/Pages/Viewer.aspx?id=ad606b10-6c80-457d-971f-686107a35636** **(https://uthvideo.uth.tmc.edu/Panopto/Pages/Viewer.aspx?id=ad606b10-6c80-457d-971f-686107a35636)**

Submit:

1) One UML Activity diagram for data acquisition, formatting and output - Total 10 points .

- Must generate UML diagram using a digital tool. No hand drawn diagrams.
- Diagram must be very specific - mention how you will process every data element

2) One ipython notebook to solve the question, and must include description of the code. {Your Python code must follow python style guide - } - 10 points

PYTHON RESTRICTION:

1. You can only import the following modules;
   1. os
   2. string
   3. csv
   4. json

   5. xml
   6. random
   7. numpy
   8. pandas
   9. scipy
2. Importing any other module will result in **ZERO** points
3. If your python code does not give correct output, you will get **ZERO** points
4. Your jupyter notebook must contain at least a) Title of project, b) Description of overall strategy, c) At least one description of important code logic just before the code, and d) title/description for output. This is in-addition to in-line comments and function docstrings. Your in-line comments must be reasonable. Do not comment every line. And yoru docstrings must be meaningful and not be full descriptions. Use Jupyter text instead. If any of the 4 components are missing, you will get **ZERO** points.

Note: Late submissions without prior permission will result in penalty (10% of max points)