Unit 10 - Homework - Data Wrangling from UMLS RRF Files

Due Nov 13, 2017 by 8am **Points** 10 **Submitting** a file upload **File Types** 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 subset of UMLS data files, located at /opt/class/umls in the class VM

From UMLS data, extract information about the following drug concepts:

1. C0016860
2. C0012265

Construct a python algorithm to extract the following information and save in a csv format.

Concept ID, Medication Name, Drug Class, Diagnosis, Mechanism of Action

Read the documentation of UMLS RRF files: **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)**

HINT:

In MRREL.RRF terms to looks for

- Drug Class = "member_of" or "CHD" or "PAR" or "isa" or "inverse_isa" as relationship type
- Diagnosis = "may_be_treated_by"
- Mechanism of Action = "mechanism_of_action_of"

There may be more than one matching result in MRREL.RRF file. If a concept is not present in MRCONSO.RRF data, you can ignore it. The dataset is manipulated in such a way that there are unique values.

Submit:

1) One ipython notebook file, and must follow style requirements for jupyter file. {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)