Class Project - III - Integrating large datasets (Insurance Claims and UMLS data)

Due Dec 4, 2017 by 8am **Points** 20 **Submitting** a file upload
File Types png and ipynb **Available** Nov 20, 2017 at 8am - Dec 6, 2017 at 8pm 17 days

This assignment was locked Dec 6, 2017 at 8pm.

Extract data from insurance claims and identify relationships.

Output:

Extract the information for the year **2009 ONLY** and provide a CSV output that has following headers:

1) PatientID
(Beneficiary code in claims data)
2) Total Costs of Outpatient care
(Total claim patyment amount for patient in given time period)
3) Predominat diagnosis
(The diagnosis that is billed the maximum number of times for the patient in the given time period. If more than one with same count, use your judgment to pick the one)
4) Medication that can treat the condition
(UMLS relationship that can treat the predominant diagnosis, If more than one with same count, use your judgment to pick the one)
5) Number of physicans visited
(Total number of unique physicians (NPI of provider) that cared for the patient)

Each row should contain data of one patient only. No duplicate rows with same patient ID. Each column should contain only one value.

Input:

1) **SynPUF data** - Medicare calims synthetic data. Contains information about cost of care. The data is available at

```
/opt/class/medicare/outpatient.csv
```

For documentation on the data, read Page 9 - of User Manual : **https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/Downloads/SynPUF_DUG.pdf (https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/Downloads/SynPUF_DUG.pdf)**

Beneficiary code can be considered Patient ID.

2) **UMLS data** - UMLS MRCONSO and MRREL rrf files. The data is filtered to contain only ICD9 data and its relationship. UMLS subset contains all ICD 9 codes, and their relationships

The data is availabel at

```
/opt/class/umls/icd/
```

The ICD 9 CM code in SynPUF data is without deciaml point. But the UMLS data contain deciaml point. Note that decimal point is not at standard location (like always 2 digits after deciaml). So when searching for the code in UML data you should consider this contrainst. You can be assured that there will be no two codes similar - like 245.8 and 24.58

Use 'may_treat' or 'may_be_treated_by' in the relationship to identify the medication relationships

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 . **You must use PANDAS library to load the SynPUF data.**

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
    10. re
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)