Project 2: i2b2 2014 Track 2 on Exploring Vital Signs and Medications
Due Date: April 1, 2021

Goals:

Use a patient clinical notes dataset for analysis. Learn its language nuances and (semi)structure.

Dataset:

The data is described in Kumar et al. paper titled 'Creation of a new longitudinal corpus of clinical narratives' which you can access from our syllabus page. The dataset which is in XML format is a mix of correspondences between medical professionals and discharge summaries after a patient visit with a doctor.

Do consult the annotation guidelines pdf file that is a part of the dataset given to you. The file is called i2b2_2014_annotation_guidelines_distribution.pdf.

Tasks: You are to conduct two types of analyses. (1) across the dataset as a whole at the note level - which means you ignore the fact that a given patient has multiple records and (2) across the dataset as a whole at the patient level – which means you do consider the fact that a given patient has multiple records.

Your analysis should be around two categories of information: (1) vital signs/physical exam readings and (2) medications. You may extract medications from the xml tagged portion at the end of each file. Note that the xml includes both specific medications as well as categories of medications. For vital signs you will need to get these from the free-text portions.

The following are questions to answer/aspects to address.

1) Provide a frequency distribution of the vital signs

2) Provide a frequency distribution of the medications taken and a frequency distribution of categories of medications taken.

3) Identify the 10 individuals taking the greatest number of medication types (this is a count of the number of medication types summed over all the records for a patient).

4) Identify the 10 individuals taking the least number of medication types (this is a count of the number of different medication types summed over all the records for a patient).

5) Identify the 10 individuals taking the least number of medications (this is a count of the number of medications summed over all the records for a patient).

6) Explore two other questions/aspects of your choice concerning vital signs and/or medications.

What you will learn at a minimum:

Working with records that have fairly complex structure.
Working with both semi-structured and unstructured free-text data.
Working on a health data analytics problem that is currently of interest to the health/medical informatics community.
Learning to use a variety of tools and toolkits
Learning to build a pipeline system from off the shelf tools to solve an important problem.
Learning to test and improve your system iteratively.
Analysis of your methods and presentation of your results
And….. learning to work collaboratively.

What to submit:

1) Each group will submit a project report presenting results by the due date. The report should include an introduction, present methods, results for each section with analysis, outline limitations and present conclusions.

2) Weekly: Each Tuesday, each group must submit a work distribution sheet that is signed by all members specifying what each person has done the previous week and will accomplish in the following week. Since we meet on zoom, one member of each group can email me the group's work distribution sheet while copying the other members on the email. I will take this to mean that group members agree with the mailed work distribution sheet. If not, let me know.

3) Final work distribution sheet: A cover sheet must be included with the final project report that summarizes the specific contributions of each member of the group towards the project.