Project 3: Exploring PUBMED data.
Due Date: May 6, 2021

Goals:

Use a PUBMED data for analysis. Learn to work with free-text of a different kind

Dataset:

Two datasets will be given to you. (1) A dataset of sentences annotated as speculative or non speculative partitioned into training and testing sets. (2) A dataset of metadata corresponding to COVID19 publications as of April 10, 2020. This dataset has several fields most of which you may ignore and focus essentially on the pubmed_id, title and abstract fields. Note that some documents do not have abstracts.

The following are your tasks with the two datasets.

A) Use the Speculative dataset to build a classifier that classifies a sentence as speculative or not. Use the train portion to build the model and use the test portion to generate results. You may explore traditional machine learning or modern deep learning approaches. Consult the demo on this given earlier. Your aim will be to get the best possible F1 score. Report F1 scores on the test set for your classifiers.

B) Analyze the COVID19 dataset in the following ways.

1) Identify the topics covered. You may use LDA for this. There are three different options for this: use the titles alone, use title + abstract and use the title + 'conclusion' sentence alone. Of the three the most interesting is to use the title + conclusion sentence. We will discuss the notion of 'conclusion sentence' in class. The optimal number of topics is likely between 5 and 10. Describe what each topic represents to the best you can and present the top 10 terms for each topic.

2) Extract and analyze the drugs and diseases mentioned in the COVID19 dataset.

3) Analyze the COVID19 dataset in **two** other ways of your choice. Get my approval on these.

What you will learn at a minimum:

Working with semi-structured / free-text records.
Exploring state of the art tools for classification and named entity recognition.
Working on a health data analytics problem that is currently of interest to the health/medical informatics community.
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.