6.S897/HST.S56 Problem Set 3
Released: Tues Mar 05
Due: Tues Mar 12 by 11:59pm through Stellar

Instructions

You will find a *pset3materials* directory in the class Github. This contains the write-up, starter notebook, and the code used to extract the data. The data is available on Physionet, specifically adult_dc_concepts.csv and adult_dc_summaries.csv . Please let us know ASAP if you have any issues accessing the data.

In the body of this exercise, blue text describes what specifically you should be submitting. If you have any questions about what the wording of the questions mean, ask us on Piazza!

For your final assessment, you should a standalone report of your analyses with relevant plots, interpretation of findings, and recommendations named ${mit_user}.pdf (e.g. *iychen.pdf*) and any code used for the write-up.

In your write-up, please make your answers as easy to identify as possible (e.g. highlight with colors, label questions with numbers, etc). The faster we can identify your answers, the faster we can grade 70 submissions! :)

Part 1: What are clinical concepts anyway? Do they work?

We are interested in clinical concept extraction, specifically recognizing UMLS terms using Metamap.

The clinical concepts extracted are located in adult_dc_concepts.csv, which are based off discharge summaries in adult_dc_summaries.csv.

1.1 How many patients have clinical concepts extracted? On average how many concepts are there per patient? Plot a histogram of clinical concepts per patient using 40 bins.

1.2 Sometimes the clinical concepts are not extracted properly. Review patient icustay_id=232593. If you search “Fruit” in the preferred_name column, what error can you spot in the extracted concepts? Why did this concept extraction mistake happen? Search through the `preferred_name` field for similar terms. What CUI number should have been used instead?

Part 2: Diseases and symptoms

We are particularly interested in diseases and symptoms.

2.1 Examine the Metamap Semantic Types. What are the symtypes that correspond to diseases and symptoms? How many unique diseases and symptoms are there in the dataset?

2.2. Examine the most frequent diseases and symptoms. What are the 5 most frequent diseases and 5 most frequent symptoms based on the clinical concept extraction?

In the notebook, we have provided some code to examine each disease / symptom and which source words were linked to that clinical concept by UMLS. As an example, we have shown that "Infantile Neuroaxonal Dystrophy" is a clinical concept tagged whenever the discharge summary mentions the word "plan", which is not what most clinicians meant when they wrote "plan" in the discharge summary.

We have provided two lists of ignore_diseases and ignore_symptoms. In the example notebook, we show how to "debug" one term by examining both the source (or "trigger") and the raw discharge summary.

2.3 Pick 3 diseases and/or symptoms from either ignore_diseases or ignore_symptoms and examine the source word. What is the trigger word? Why did we choose to remove this clinical concept? Some are more subtle than others, so you may need to search in clinical literature depending on your medical background.

2.4 Besides removing mis-tagged clinical concepts altogether, what is another data cleaning technique we might use to get better extracted diseases and symptoms?

Part 3: Building a graph of medicine

Using these extract clinical concepts, we want to understand how diseases and symptoms relate. If we remove the disease and symptoms mentioned in the ignore lists, we have a set of "clean" diseases and symptoms. Build an associative graph between disease and symptom co-occurrences for the top 100 most frequent diseases and symptoms -- removing the diseases and symptoms from the ignore lists. That is, we want a matrix M where the rows correspond to diseases and the columns correspond to symptoms. Each entry should be the count of co-occurrences for that disease and symptom: that is, for disease i and symptoms j, how many patients had both disease i and symptom j according to the extracted clinical concepts? If there are 104 patients with disease i and symptom j, M[i,j] = 104.

Additionally, we want to count the frequencies of diseases and symptoms overall. Create array D where D[i] is the number of patients with disease i, and array S where S[j] is the number of patients with symptom j.

3.1 Read Finlayson et al, 2014, a paper using similar extracted clinical concepts. What is one main difference between our approach thus far and the authors?

We are most interested in the concept of lift as defined in Finlayson et al, 2014. Let P(“Pneumonia” = 1) denote the probability that a patient has the clinical concept “Pneumonia” extracted from the discharge summary.

3.2 Given our empirical counts, how can we estimate P(“Pneumonia” = 1)? How can we estimate lift(“Dry cough” -> “Pneumonia”)? Please express your answer as as a function of M, D, S, and the number of total patients N and show your derivation.

3.3 Using our empirical numbers, calculate the lift(S -> D) for a given D and for all S in the set of symptoms. We can sort by symptoms to get the most meaningful symptoms.

- What are the 5 symptoms with the highest lift(S -> “Pneumonia”)? What is pneumonia and what are the most common symptoms? How well does this align?
- What are the 5 symptoms have the highest lift(S -> “Hypothyroidism”)? What is hypothyroidism and what are the most common symptoms? How well does this align?

Part 4: Not graded and please answer

4.1 How many hours did this problem set take you?
4.2 How many hours a week do you spend on this class (attending lecture, doing readings, working on problem sets) on average?