**Overall Plan**

Initially I thought the task goal was to develop models to distinguish between different conditions that the patients within the dataset might have. To that end, my plan was to use similarity measures to cluster the patients into groups and determine what factors might be important. My idea here was to

I realized that the task might also be that all the patients within the dataset had the same unknown condition, and the model was to determine what commonalities that they patients had.

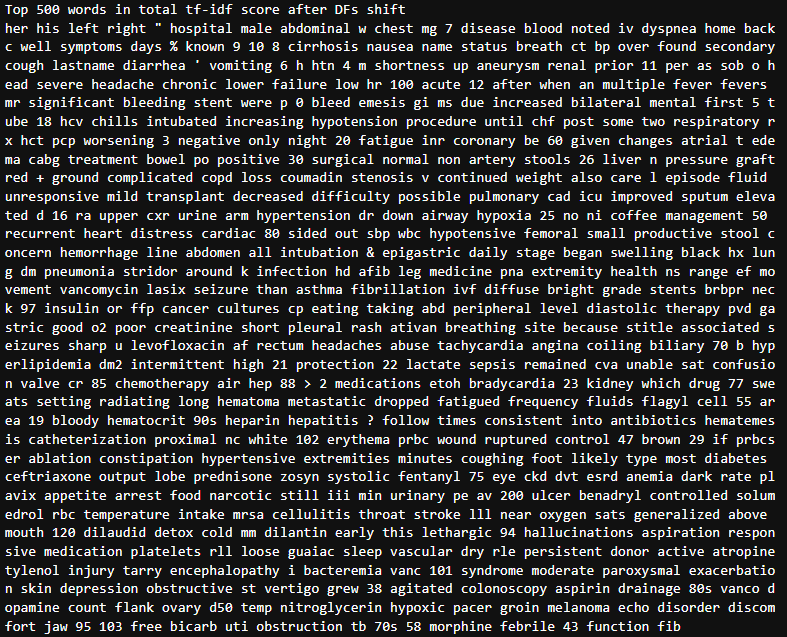
**Data Preprocessing**

From the text files and NER results given, I first extracted the Chief Complaint and History of Present Illness sections from each of the patient history files. I also wrote a function to convert a .ann into two pandas data frames, one containing each of the entities and one with the relations between the entities. (convert\_files.py > patient\_record\_summary.csv, ann\_train.txt)

**Factors**

I computed the TF-IDF values for each of the documents in the patient\_record\_summary.csv. I artificially increased the DF values for each term to manually decrease the final scores of words that were very common. (tfidfs.py > word\_frequencies.png)

Afterwards, I summed the total TF-IDF values for each term and printed the top most values. Some common medically relevant words were “abdominal, chest, blood, dyspnea, cirrhosis, nausea, breath, cough, diarrhea, vomiting, shortness,…” with many terms involving the chest and lungs.



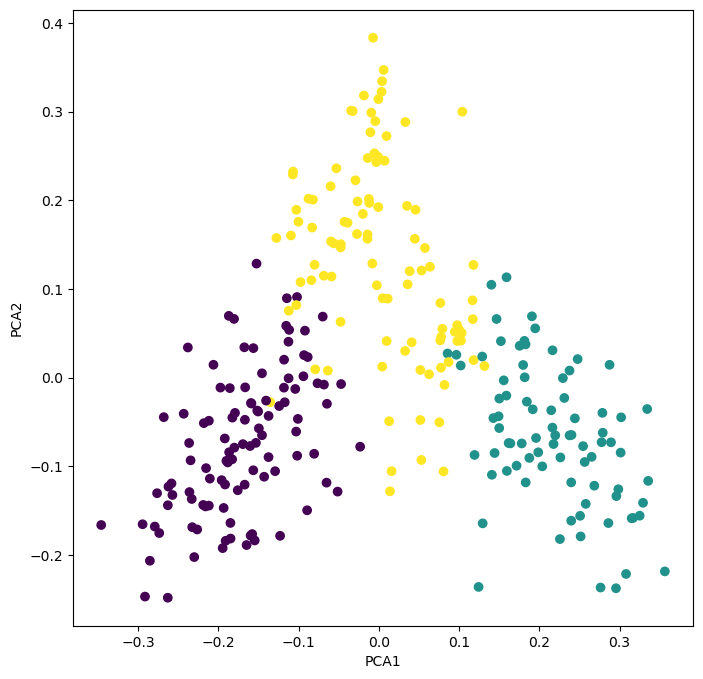
**Clustering**

To cluster the patients, I attempted to perform TF-IDF on

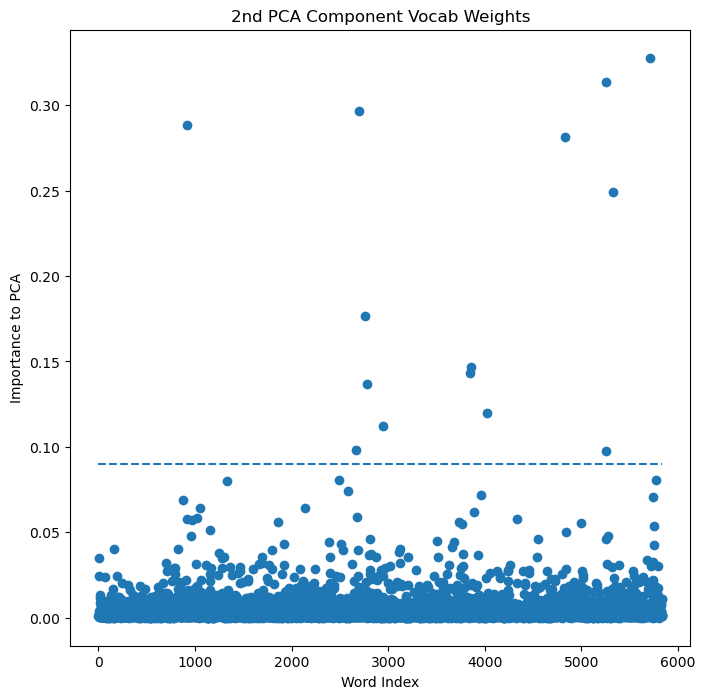
* The chief complaints and history of present illnesses
* Each of the reasons and drugs from the NER files

I also attempted to tune a language model (prajjwal1/bert-tiny) from HuggingFace on the same chief complaints/history text, and had the idea to perform a masked prediction task on Reason-Drug pairs from the NER results. The hope was that the language model might learn to predict what symptoms a patient might present based on their other symptoms, or what drug would be used to treat what symptom, etc. With this model we could encode the reasons and drugs from the NER model and use this for similarity analysis, but I had difficulties with the model. (train\_lm.py > text\_model, text\_tokenizer)

Here I used sklearn to compute the TF-IDF vectors, using the learned vectors for PCA and K-means clustering (3 clusters). (cluster.py > clusters.png, pca\_importance.png)



Although separation looks good, analyzing the weights of terms contributing to first two PCA directions shows that this is mainly a variation of gender.



A screenshot of a computer screen

Description automatically generated with medium confidence

Although in the 3rd PCA component, we see that “chest” and “abdominal” show up as being heavily weighted. Notably, these have the same sign in weight, showing that patients with mentions to the chest also have mentions to the abdomen.

**Further Strategies**

* With more data, we could develop a language model to extract vital statistics of the patients to perform more statistical analysis with actual numbers
* We could certainly do more with the NER data provided, especially if we could model what symptoms (Reasons) and drugs are commonly linked with what types of conditions
* More information is available in the remainder of the patient file, the richness of this data lends to language modeling having the potential to provide more insights