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.

```
Top 500 words in total tf-idf score after DFs shift
her his left right " hospital male abdominal w chest mg 7 disease blood noted iv dyspnea home back
c well symptoms days % known 9 10 8 cirrhosis nausea name status breath ct bp over found secondary
cough lastname diarrhea ' vomiting 6 h htn 4 m shortness up aneurysm renal prior 11 per as sob o h
ead severe headache chronic lower failure low hr 100 acute 12 after when an multiple fever fevers
mr significant bleeding stent were p 0 bleed emesis gi ms due increased bilateral mental first 5 t
ube 18 hcv chills intubated increasing hypotension procedure until chf post some two respiratory r
x hct pcp worsening 3 negative only night 20 fatigue inr coronary be 60 given changes atrial t ede
ma cabg treatment bowel po positive 30 surgical normal non artery stools 26 liver n pressure graft
red + ground complicated copd loss coumadin stenosis v continued weight also care l episode fluid
unresponsive mild transplant decreased difficulty possible pulmonary cad icu improved sputum eleva
ted d 16 ra upper cxr urine arm hypertension dr down airway hypoxia 25 no ni coffee management 50
recurrent heart distress cardiac 80 sided out sbp wbc hypotensive femoral small productive stool c
oncern hemorrhage line abdomen all intubation & epigastric daily stage began swelling black hx lun
g dm pneumonia stridor around k infection hd afib leg medicine pna extremity health ns range ef mo
vement vancomycin lasix seizure than asthma fibrillation ivf diffuse bright grade stents brbpr nec
k 97 insulin or ffp cancer cultures cp eating taking abd peripheral level diastolic therapy pvd ga
stric good o2 poor creatinine short pleural rash ativan breathing site because stitle associated s
eizures sharp u levofloxacin af rectum headaches abuse tachycardia angina coiling biliary 70 b hyp
erlipidemia dm2 intermittent high 21 protection 22 lactate sepsis remained cva unable sat confusio
n valve cr 85 chemotherapy air hep 88 > 2 medications etoh bradycardia 23 kidney which drug 77 swe
ats setting radiating long hematoma metastatic dropped fatigued frequency fluids flagyl cell 55 ar
ea 19 bloody hematocrit 90s heparin hepatitis ? follow times consistent into antibiotics hematemes
is catheterization proximal nc white 102 erythema prbc wound ruptured control 47 brown 29 if prbcs
er ablation constipation hypertensive extremities minutes coughing foot likely type most diabetes
ceftriaxone output lobe prednisone zosyn systolic fentanyl 75 eye ckd dvt esrd anemia dark rate pl
avix appetite arrest food narcotic still iii min urinary pe av 200 ulcer benadryl controlled solum
edrol rbc temperature intake mrsa cellulitis throat stroke lll near oxygen sats generalized above
mouth 120 dilaudid detox cold mm dilantin early this lethargic 94 hallucinations aspiration respon
sive medication platelets rll loose guaiac sleep vascular dry rle persistent donor active atropine
tylenol injury tarry encephalopathy i bacteremia vanc 101 syndrome moderate paroxysmal exacerbatio
n skin depression obstructive st vertigo grew 38 agitated colonoscopy aspirin drainage 80s vanco d
opamine count flank ovary d50 temp nitroglycerin hypoxic pacer groin melanoma echo disorder discom
fort jaw 95 103 free bicarb uti obstruction tb 70s 58 morphine febrile 43 function fib
```

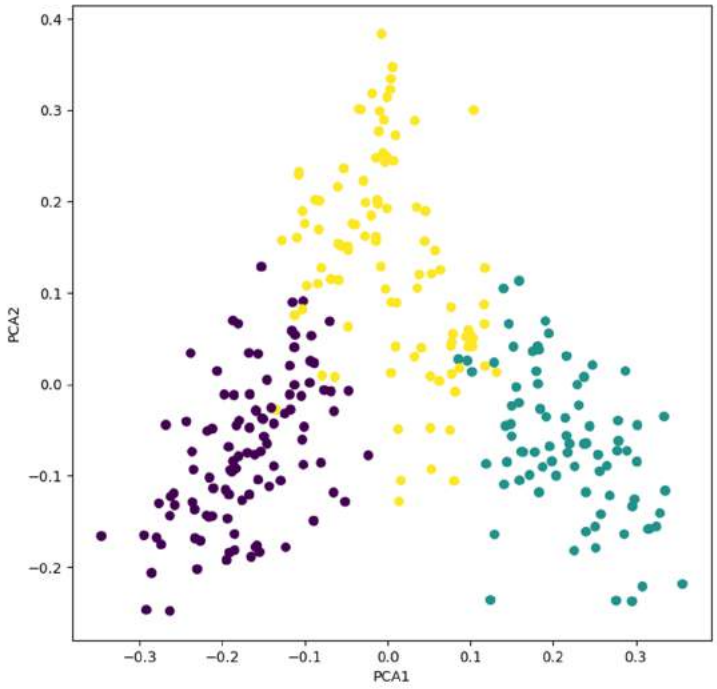
Clustering

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

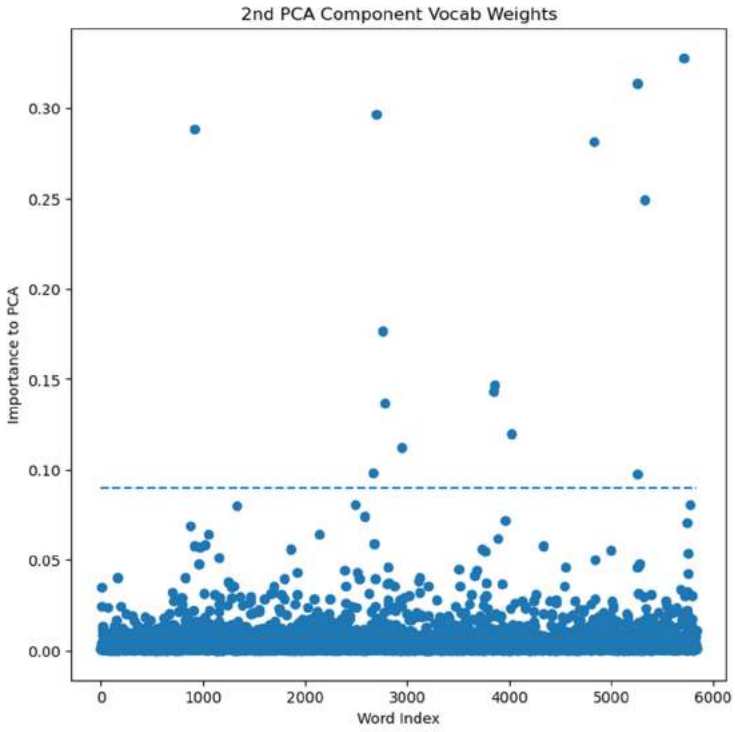
- The chief complaints and history of present illnesses
- Each of the reasons and drugs from the NER files

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.



```
Words important for the 1th pca component
['he', 'her', 'his', 'she']
Words important for the 2th pca component
['and', 'had', 'he', 'her', 'his', 'in', 'of', 'on', 'patient', 'she', 'that', 'the', 'to', 'was']
Words important for the 3th pca component
['abdominal', 'and', 'chest', 'denies', 'has', 'he', 'no', 'on', 'or', 'pain', 'patient', 'reports', 'she', 'the', 'to',
 'was']
Words important for the 4th pca component
['and', 'on', 'patient', 'pt', 'the', 'was']
Words important for the 5th pca component
['and', 'cath', 'chest', 'denies', 'he', 'her', 'his', 'of', 'pain', 'patient', 'pt', 'she', 'with']
```

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