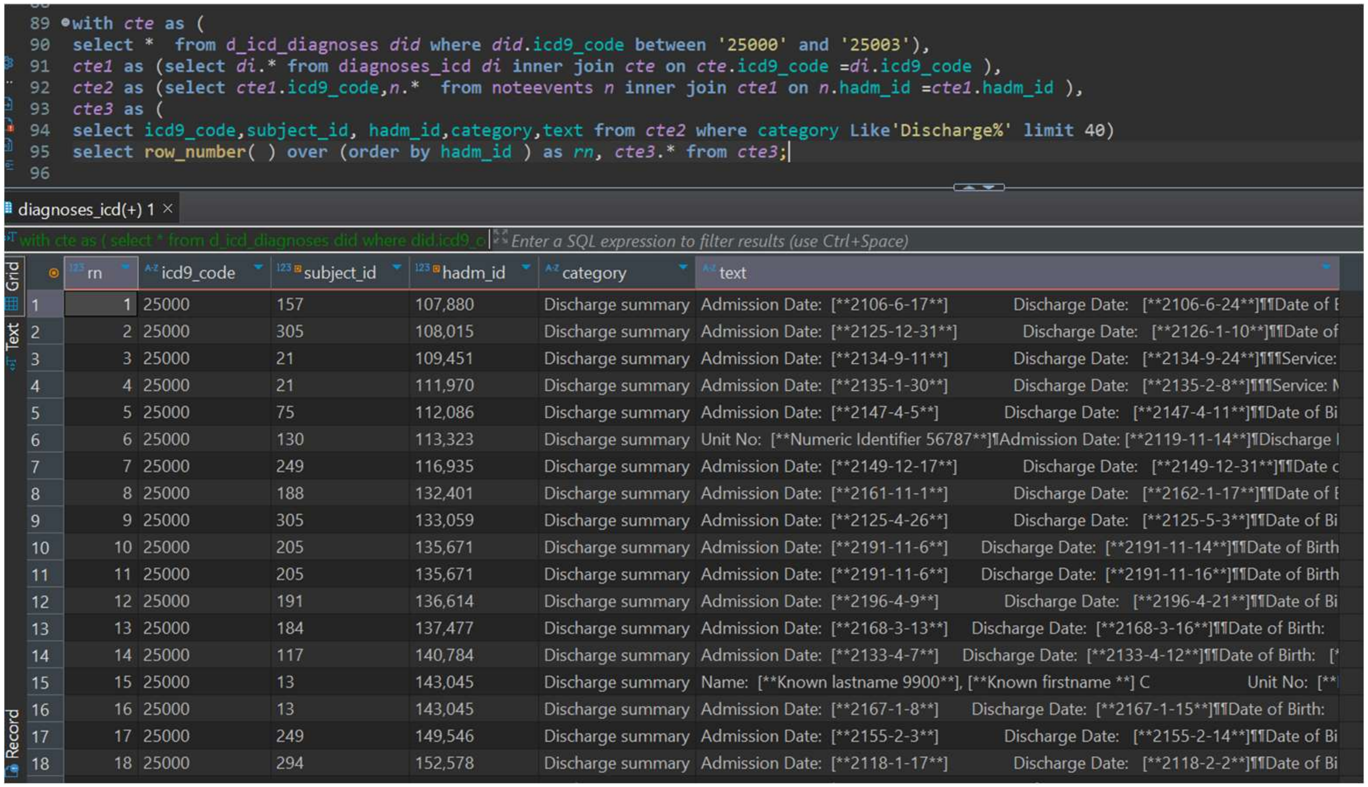
MIMIC NLP

Tools and Dataset Used:

- Filename: https://github.com/dalwari/mimic-iii-clinical-database-demo-1.4/blob/main/mimic_nlp.ipynb
- Dataset Used: https://github.com/dalwari/mimic-iii-clinical-database-demo-1.4/blob/main/nlp_med_notes.csv
- Package Used: Spacy, Scispacy, Word2Vec, TSNE
- Application Used: Dbeaver, VSCode

Dataset Preparation:

- Write a sql query in Dbeaver
- Use the ‘Export data’ option in Dbeaver to extract the dataset result as CSV file into NLP workspace

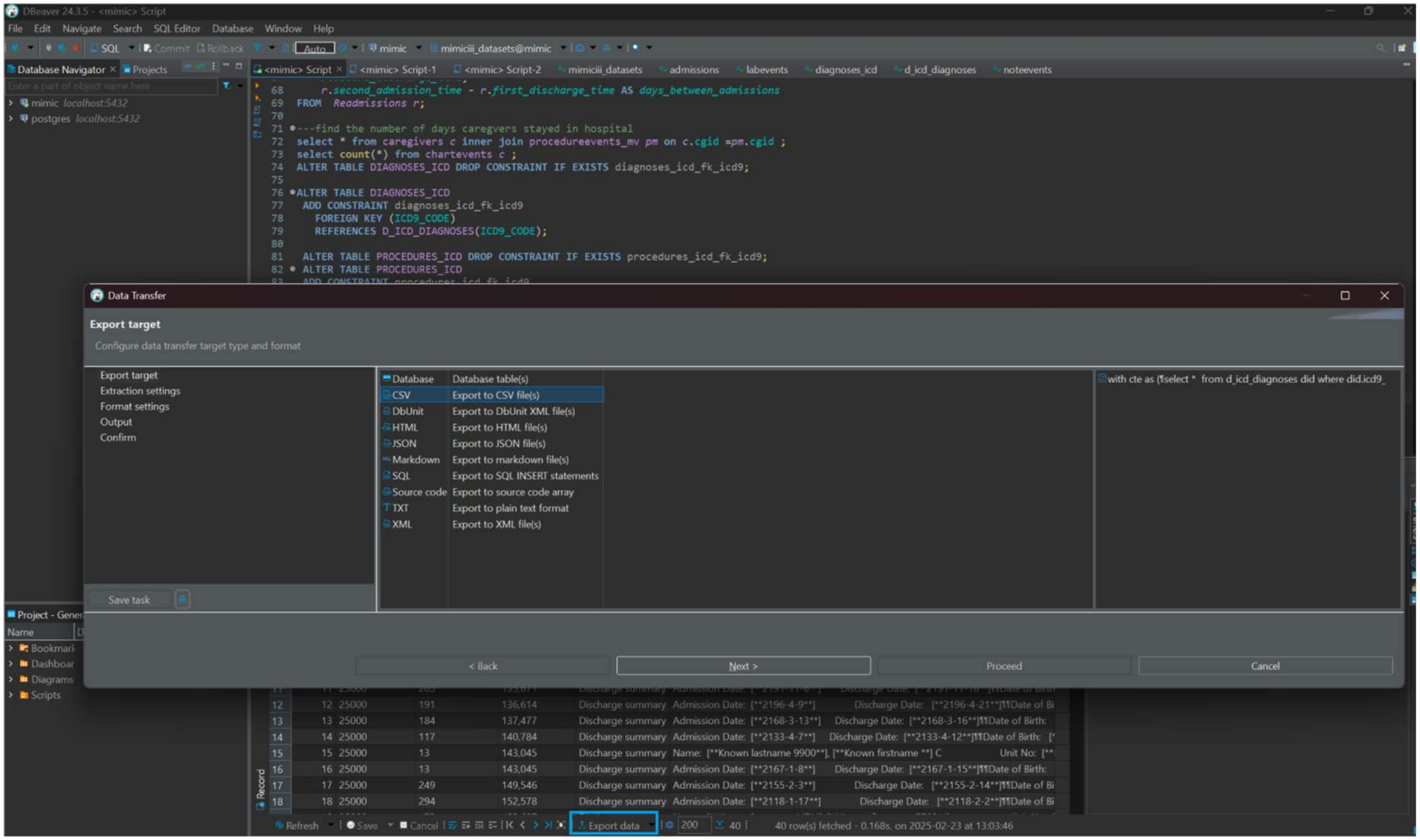


```sql
with cte as (
select *  from d_icd_diagnoses did where did.icd9_code between '25000' and '25003'),
cte1 as (select di.* from diagnoses_icd di inner join cte on cte.icd9_code =di.icd9_code ),
cte2 as (select cte1.icd9_code,n.*  from noteevents n inner join cte1 on n.hadm_id =cte1.hadm_id ),
cte3 as (
select icd9_code,subject_id, hadm_id,category,text from cte2 where category Like'Discharge%' limit 40)
select row_number( ) over (order by hadm_id ) as rn, cte3.* from cte3;
```

diagnoses_icd(+) 1
with cte as (select * from d_icd_diagnoses did where did.icd9_c
Enter a SQL expression to filter results (use Ctrl+Space)
Grid
Text
Record

	rn	icd9_code	subject_id	hadm_id	category	text
1	1	25000	157	107,880	Discharge summary	Admission Date: [**2106-6-17**] Discharge Date: [**2106-6-24**]¶¶Date of
2	2	25000	305	108,015	Discharge summary	Admission Date: [**2125-12-31**] Discharge Date: [**2126-1-10**]¶¶Date of
3	3	25000	21	109,451	Discharge summary	Admission Date: [**2134-9-11**] Discharge Date: [**2134-9-24**]¶¶¶Service:
4	4	25000	21	111,970	Discharge summary	Admission Date: [**2135-1-30**] Discharge Date: [**2135-2-8**]¶¶¶Service:
5	5	25000	75	112,086	Discharge summary	Admission Date: [**2147-4-5**] Discharge Date: [**2147-4-11**]¶¶Date of Bi
6	6	25000	130	113,323	Discharge summary	Unit No: [**Numeric Identifier 56787**]¶Admission Date: [**2119-11-14**]¶Discharge
7	7	25000	249	116,935	Discharge summary	Admission Date: [**2149-12-17**] Discharge Date: [**2149-12-31**]¶¶Date
8	8	25000	188	132,401	Discharge summary	Admission Date: [**2161-11-1**] Discharge Date: [**2162-1-17**]¶¶Date of
9	9	25000	305	133,059	Discharge summary	Admission Date: [**2125-4-26**] Discharge Date: [**2125-5-3**]¶¶Date of Bi
10	10	25000	205	135,671	Discharge summary	Admission Date: [**2191-11-6**] Discharge Date: [**2191-11-14**]¶¶Date of Birth
11	11	25000	205	135,671	Discharge summary	Admission Date: [**2191-11-6**] Discharge Date: [**2191-11-16**]¶¶Date of Birth
12	12	25000	191	136,614	Discharge summary	Admission Date: [**2196-4-9**] Discharge Date: [**2196-4-21**]¶¶Date of Bi
13	13	25000	184	137,477	Discharge summary	Admission Date: [**2168-3-13**] Discharge Date: [**2168-3-16**]¶¶Date of Birth:
14	14	25000	117	140,784	Discharge summary	Admission Date: [**2133-4-7**] Discharge Date: [**2133-4-12**]¶¶Date of Birth: [
15	15	25000	13	143,045	Discharge summary	Name: [**Known lastname 9900**], [**Known firstname **] C Unit No: [**
16	16	25000	13	143,045	Discharge summary	Admission Date: [**2167-1-8**] Discharge Date: [**2167-1-15**]¶¶Date of Birth:
17	17	25000	249	149,546	Discharge summary	Admission Date: [**2155-2-3**] Discharge Date: [**2155-2-14**]¶¶Date of Bi
18	18	25000	294	152,578	Discharge summary	Admission Date: [**2118-1-17**] Discharge Date: [**2118-2-2**]¶¶Date of Bi

- The disease of interest here is diabetes
- The icd9_code for diabetes is between '25000' and '25003'
- First we need to join d_icd_diagnoses and diagnoses_icd
- The result of previous step subquery is joined with noteevents on hospital admission id.
- The notes of 'Discharge Summary' needs to be filtered out limited to 40 records in order to process effectively.


DBeaver 24.3.5 - <mimic> Script
File Edit Navigate Search SQL Editor Database Window Help
SQL Commit Rollback Auto mimic mimiciii_datasets@mimic
Database Navigator Projects
mimic localhost:5432
postgres localhost:5432
<mimic> Script <mimic> Script-1 <mimic> Script-2 mimiciii_datasets admissions labevents diagnoses_icd d_icd_diagnoses noteevents

```
    r.second_admission_time - r.first_discharge_time AS days_between_admissions
FROM  Readmissions r;

---find the number of days caregvers stayed in hospital
select * from caregivers c inner join procedureevents_mv pm on c.cgid =pm.cgid ;
select count(*) from chartevents c ;
ALTER TABLE DIAGNOSES_ICD DROP CONSTRAINT IF EXISTS diagnoses_icd_fk_icd9;

ALTER TABLE DIAGNOSES_ICD
 ADD CONSTRAINT diagnoses_icd_fk_icd9
   FOREIGN KEY (ICD9_CODE)
   REFERENCES D_ICD_DIAGNOSES(ICD9_CODE);

 ALTER TABLE PROCEDURES_ICD DROP CONSTRAINT IF EXISTS procedures_icd_fk_icd9;
 ALTER TABLE PROCEDURES_ICD
```

Data Transfer

Export target

Configure data transfer target type and format

Database	Database table(s)
CSV	Export to CSV file(s)
DbUnit	Export to DbUnit XML file(s)
HTML	Export to HTML file(s)
JSON	Export to JSON file(s)
Markdown	Export to markdown file(s)
SQL	Export to SQL INSERT statements
Source code	Export to source code array
TXT	Export to plain text format
XML	Export to XML file(s)

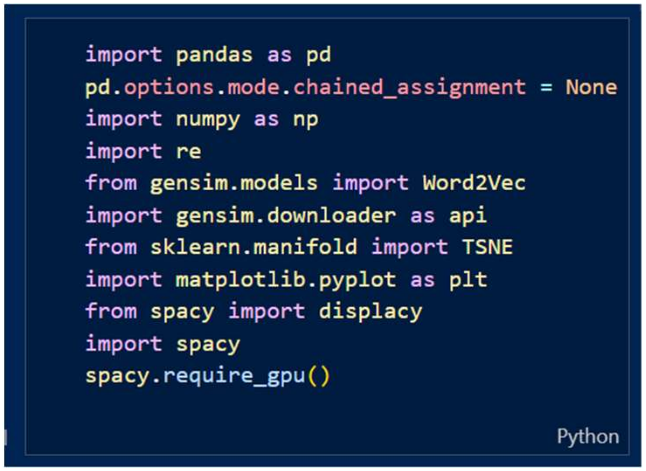
with cte as (¶select * from d_icd_diagnoses did where did.icd9_

12	12 25000	191	136,614	Discharge summary	Admission Date: [**2196-4-9**] Discharge Date: [**2196-4-21**]¶¶Date of Bi
13	13 25000	184	137,477	Discharge summary	Admission Date: [**2168-3-13**] Discharge Date: [**2168-3-16**]¶¶Date of Birth:
14	14 25000	117	140,784	Discharge summary	Admission Date: [**2133-4-7**] Discharge Date: [**2133-4-12**]¶¶Date of Birth: [
15	15 25000	13	143,045	Discharge summary	Name: [**Known lastname 9900**], [**Known firstname **] C Unit No: [**
16	16 25000	13	143,045	Discharge summary	Admission Date: [**2167-1-8**] Discharge Date: [**2167-1-15**]¶¶Date of Birth:
17	17 25000	249	149,546	Discharge summary	Admission Date: [**2155-2-3**] Discharge Date: [**2155-2-14**]¶¶Date of Bi
18	18 25000	294	152,578	Discharge summary	Admission Date: [**2118-1-17**] Discharge Date: [**2118-2-2**]¶¶Date of Bi

40 row(s) fetched - 0.168s, on 2025-02-23 at 13:03:46

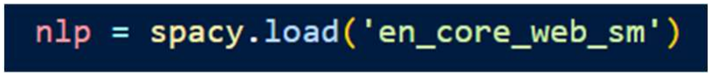
Named Entity Recognition:

- Import the Spacy, Scispacy, Word2Vec, TSNE libraries



```python
import pandas as pd
pd.options.mode.chained_assignment = None
import numpy as np
import re
from gensim.models import Word2Vec
import gensim.downloader as api
from sklearn.manifold import TSNE
import matplotlib.pyplot as plt
from spacy import displacy
import spacy
spacy.require_gpu()
```

- Load the spacy model



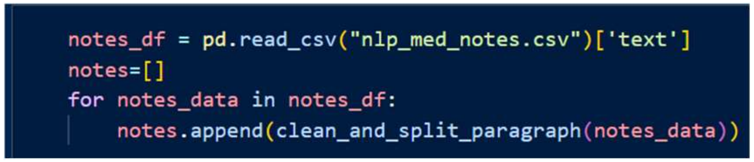
```python
nlp = spacy.load('en_core_web_sm')
```

- Prepare utility functions like,
- clean_and_split_paragraph()
- extract_entities()
- fetch_entities()
- visualize_entities()
- extract_corpus()
- fetch_corpus()
- extract_passage_by_label()
- do_data_process()

Utility Helper Functions:

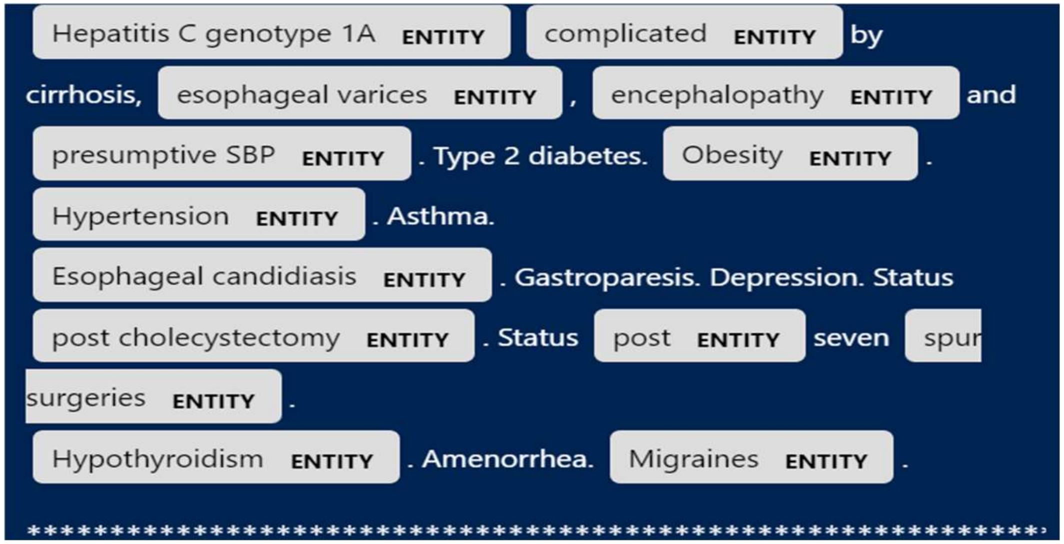
- **clean_and_split_paragraph()** is used to perform data cleaning by removing the extra spaces and redundant lines.
- **fetch_entities()** and **extract_entities()** are used to print entities fetched using NER models.
- **fetch_corpus()** and **extract_corpus()** are used to print corpus recognized from the dataset.
- **do_data_process()** and **extract_passage_by_label()** are used to extract subparagraph from medical notes passage.
- **visualize_entities()** is used to identitfy the different types of entities provided by different *Spacy* and *SciSpacy* models
- For instance , *History of Present Illness, Past Medical History, Brief Hospital Course, REVIEW OF SYSTEMS* are the subparagraph labels

- Load and clean the ‘nlp_med_notes.csv’ dataset

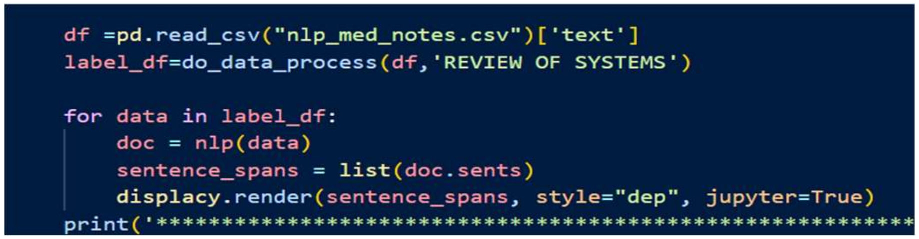


```
notes_df = pd.read_csv("nlp_med_notes.csv")['text']
notes=[]
for notes_data in notes_df:
    notes.append(clean_and_split_paragraph(notes_data))
```

- For each record extract the ‘PAST MEDICAL HISTORY’ subparagraph from text (i.e medical notes) field.
- Visualize the extracted label_df to view the entities of spacy model.



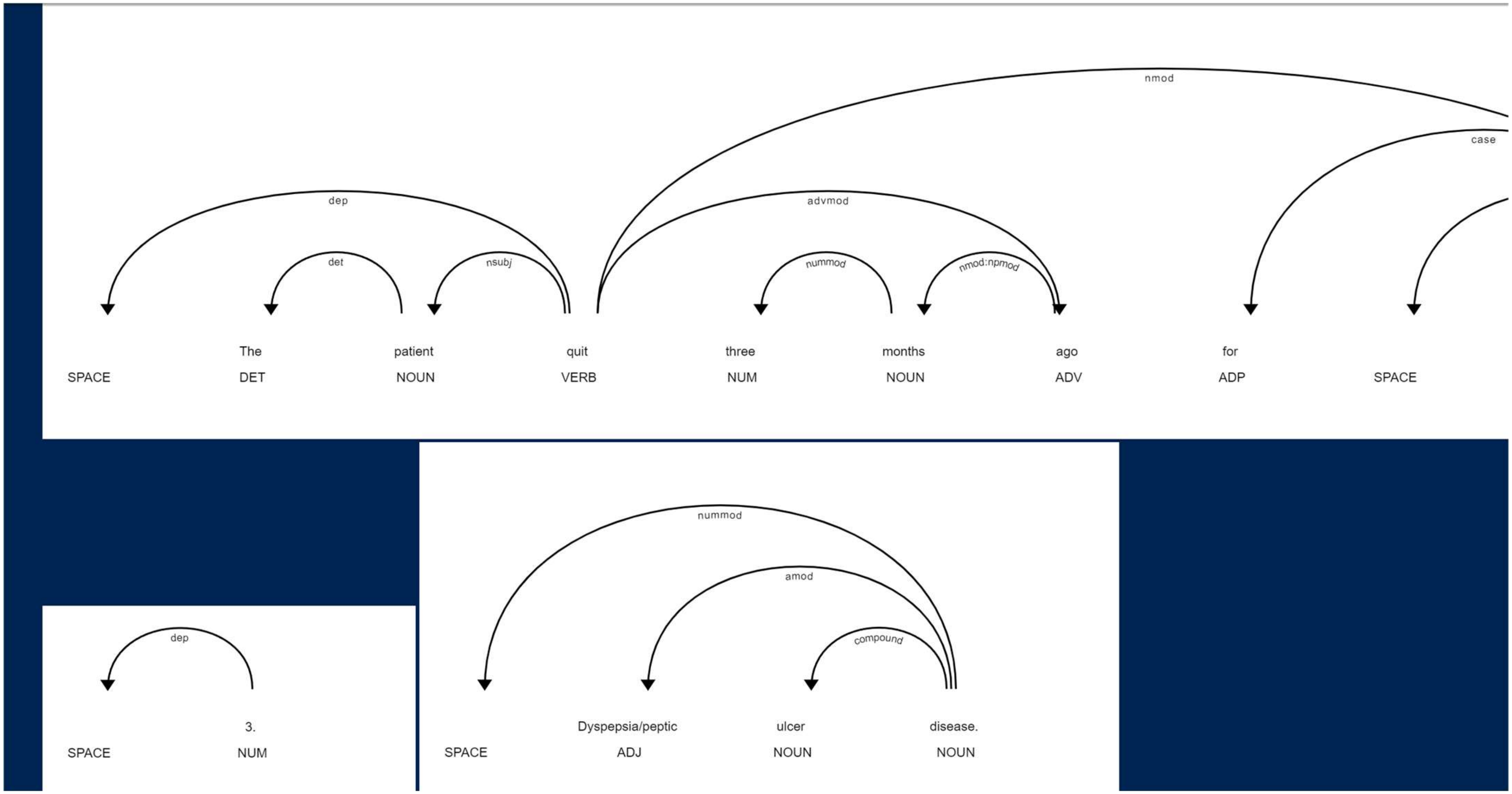
Hepatitis C genotype 1A ENTITY complicated ENTITY by
cirrhosis, esophageal varices ENTITY , encephalopathy ENTITY and
presumptive SBP ENTITY . Type 2 diabetes. Obesity ENTITY .
Hypertension ENTITY . Asthma.
Esophageal candidiasis ENTITY . Gastroparesis. Depression. Status
post cholecystectomy ENTITY . Status post ENTITY seven spur
surgeries ENTITY .
Hypothyroidism ENTITY . Amenorrhea. Migraines ENTITY .
****

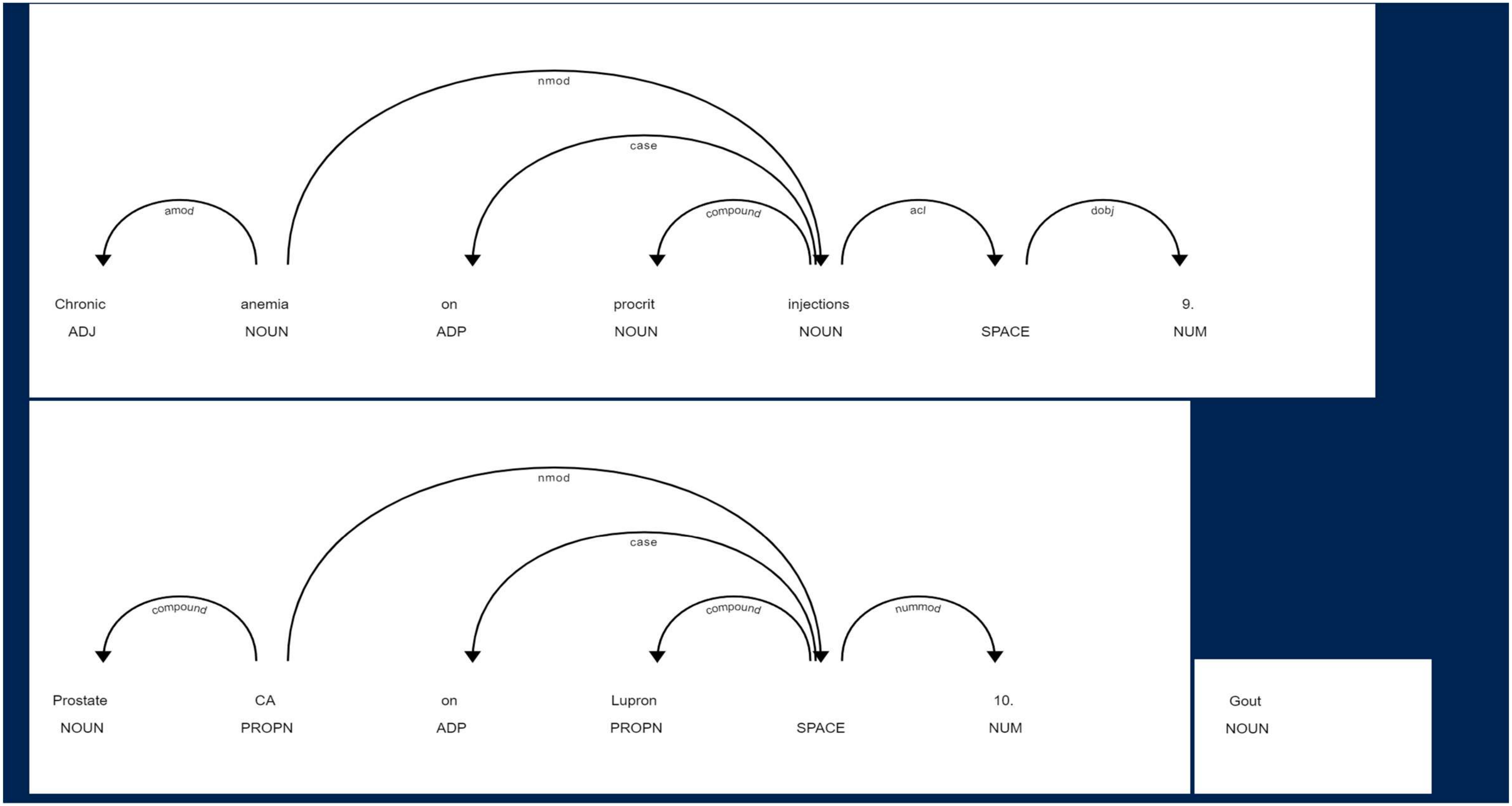


```
df =pd.read_csv("nlp_med_notes.csv")['text']
label_df=do_data_process(df,'REVIEW OF SYSTEMS')

for data in label_df:
    doc = nlp(data)
    sentence_spans = list(doc.sents)
    displacy.render(sentence_spans, style="dep", jupyter=True)
print('**********************************************************
```

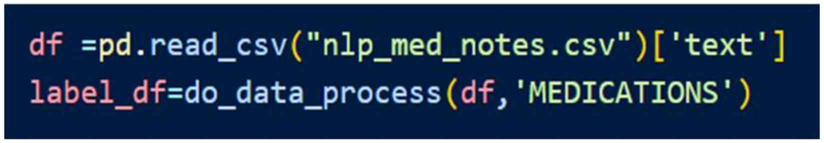
- Create a **dependency tree** using displacy from spacy model to show the relationship between words in a sentence.
- A meaningful sentence can be extracted from ‘REVIEW OF SYSTEMS’ sub paragraph.


nmod
case
dep
advmod
det
nsubj
nummod
nmod:npmod
SPACE
The DET
patient NOUN
quit VERB
three NUM
months NOUN
ago ADV
for ADP
SPACE
dep
SPACE
3. NUM
nummod
amod
compound
SPACE
Dyspepsia/peptic ADJ
ulcer NOUN
disease. NOUN

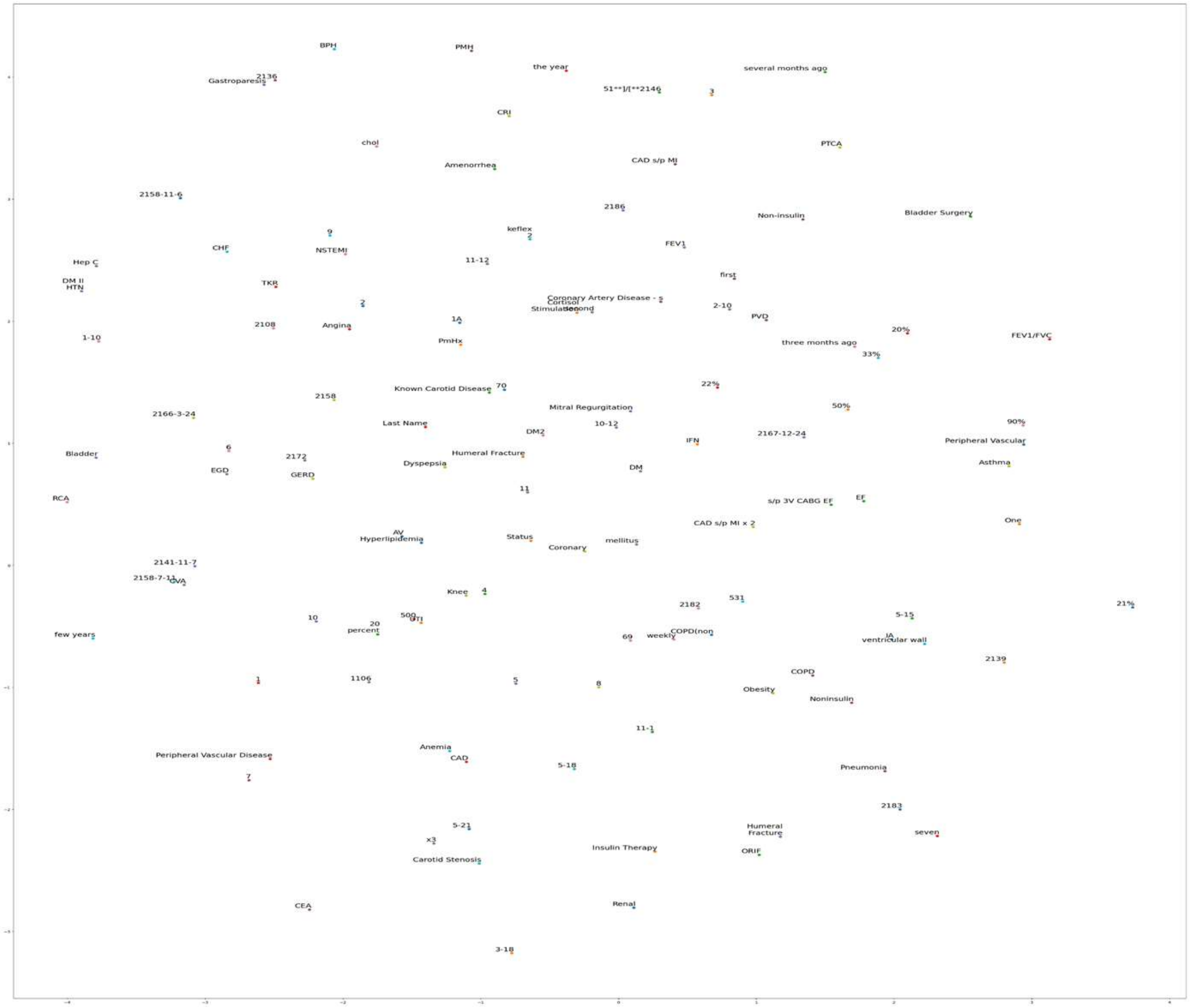

nmod
case
amod
compound
acl
dobj
Chronic
ADJ
anemia
NOUN
on
ADP
procrit
NOUN
injections
NOUN
SPACE
9.
NUM
nmod
case
compound
compound
nummod
Prostate
NOUN
CA
PROPN
on
ADP
Lupron
PROPN
SPACE
10.
NUM
Gout
NOUN

Spacy TSNE visualization: For Medication in medical notes

https://github.com/dalwari/mimic-iii-clinical-database-demo-1.4/blob/main/spacy_tsne_plot_diag.png



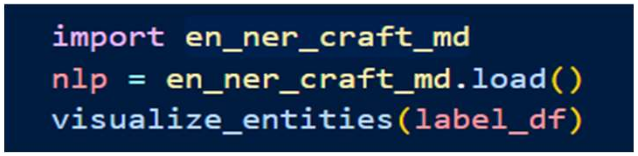
```
df =pd.read_csv("nlp_med_notes.csv")['text']
label_df=do_data_process(df,'MEDICATIONS')
```



SciSpacy Usage:

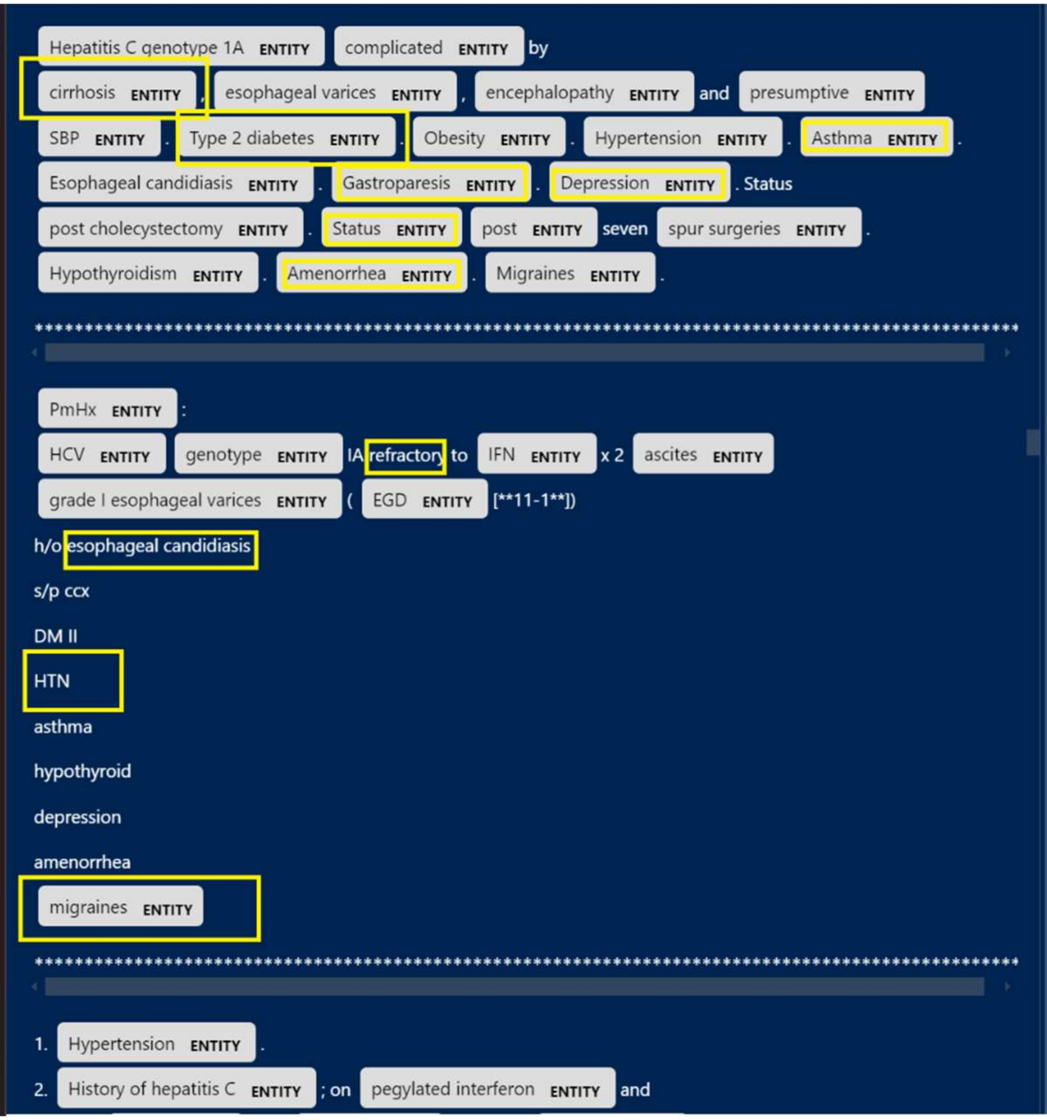
- SciSpacy models:
- en_core_sci_md
- en_core_sci_lg
- en_ner_craft_md
- en_ner_jnlpba_md
- en_ner_bionlp13cg_md
- en_ner_bc5cdr_md

- Load the model and visualize the corpus.



```
import en_ner_craft_md
nlp = en_ner_craft_md.load()
visualize_entities(label_df)
```

Sciscpacy: ***en_core_sci_lg***



Hepatitis C genotype 1A ENTITY complicated ENTITY by

cirrhosis ENTITY , esophageal varices ENTITY , encephalopathy ENTITY and presumptive ENTITY

SBP ENTITY . Type 2 diabetes ENTITY . Obesity ENTITY . Hypertension ENTITY . Asthma ENTITY .

Esophageal candidiasis ENTITY . Gastroparesis ENTITY . Depression ENTITY . Status

post cholecystectomy ENTITY . Status ENTITY post ENTITY seven spur surgeries ENTITY .

Hypothyroidism ENTITY . Amenorrhea ENTITY . Migraines ENTITY .

****

PmHx ENTITY :

HCV ENTITY genotype ENTITY IA refractory to IFN ENTITY x 2 ascites ENTITY

grade I esophageal varices ENTITY (EGD ENTITY [**11-1**])

h/o esophageal candidiasis

s/p ccx

DM II

HTN

asthma

hypothyroid

depression

amenorrhea

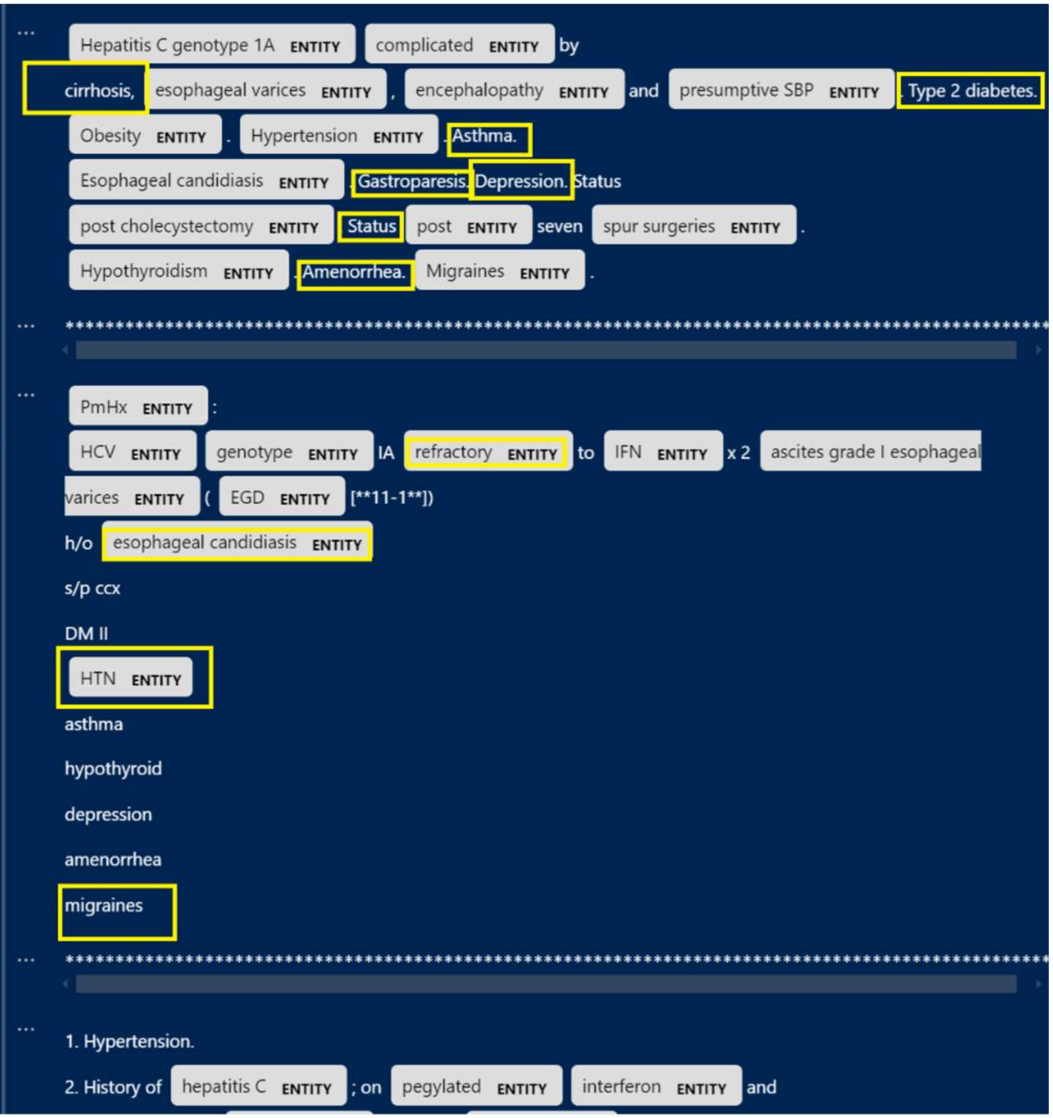
migraines ENTITY

****

1. Hypertension ENTITY .

2. History of hepatitis C ENTITY ; on pegylated interferon ENTITY and

Spacy:



...

Hepatitis C genotype 1A ENTITY complicated ENTITY by

cirrhosis, esophageal varices ENTITY , encephalopathy ENTITY and presumptive SBP ENTITY . Type 2 diabetes.

Obesity ENTITY . Hypertension ENTITY . Asthma.

Esophageal candidiasis ENTITY . Gastroparesis. Depression. Status

post cholecystectomy ENTITY . Status post ENTITY seven spur surgeries ENTITY .

Hypothyroidism ENTITY . Amenorrhea. Migraines ENTITY .

...

****

...

PmHx ENTITY :

HCV ENTITY genotype ENTITY IA refractory ENTITY to IFN ENTITY x 2 ascites grade I esophageal

varices ENTITY (EGD ENTITY [**11-1**])

h/o esophageal candidiasis ENTITY

s/p ccx

DM II

HTN ENTITY

asthma

hypothyroid

depression

amenorrhea

migraines

...

****

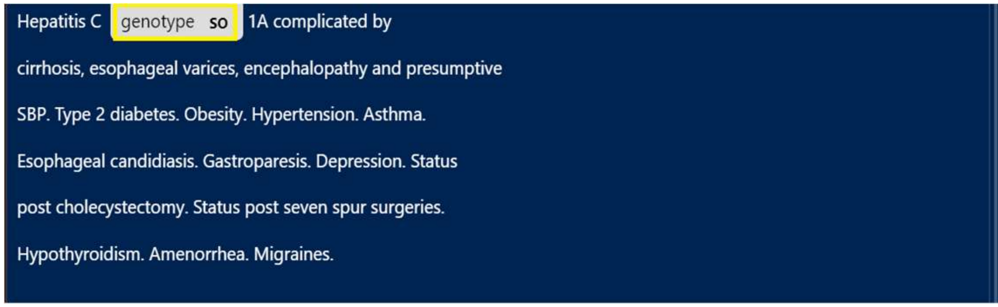
...

1. Hypertension.

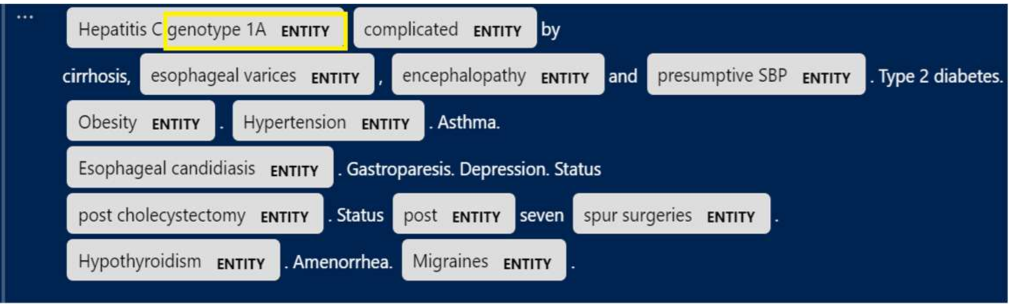
2. History of hepatitis C ENTITY ; on pegylated ENTITY interferon ENTITY and

Scipacy: ***en_ner_craft_md***

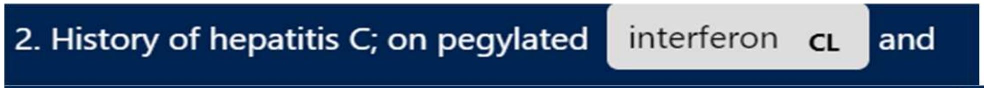
Spacy:



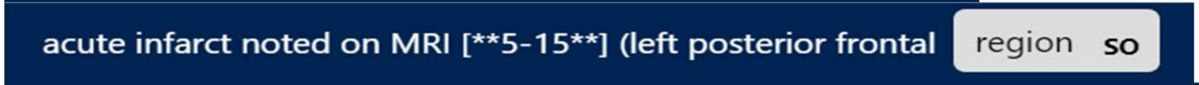
Hepatitis C genotype SO 1A complicated by cirrhosis, esophageal varices, encephalopathy and presumptive SBP. Type 2 diabetes. Obesity. Hypertension. Asthma. Esophageal candidiasis. Gastroparesis. Depression. Status post cholecystectomy. Status post seven spur surgeries. Hypothyroidism. Amenorrhea. Migraines.



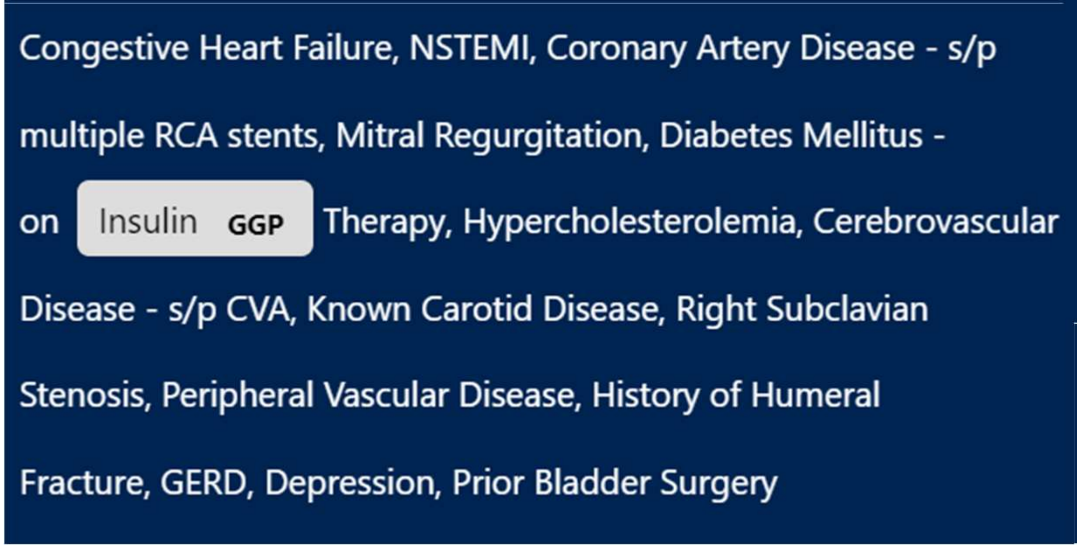
Hepatitis C genotype 1A ENTITY complicated ENTITY by cirrhosis, esophageal varices ENTITY , encephalopathy ENTITY and presumptive SBP ENTITY . Type 2 diabetes. Obesity ENTITY . Hypertension ENTITY . Asthma. Esophageal candidiasis ENTITY . Gastroparesis. Depression. Status post cholecystectomy ENTITY . Status post ENTITY seven spur surgeries ENTITY . Hypothyroidism ENTITY . Amenorrhea. Migraines ENTITY .



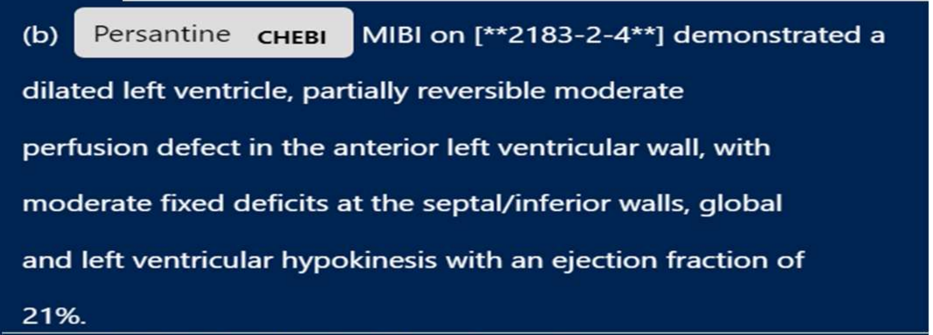
2. History of hepatitis C; on pegylated interferon CL and



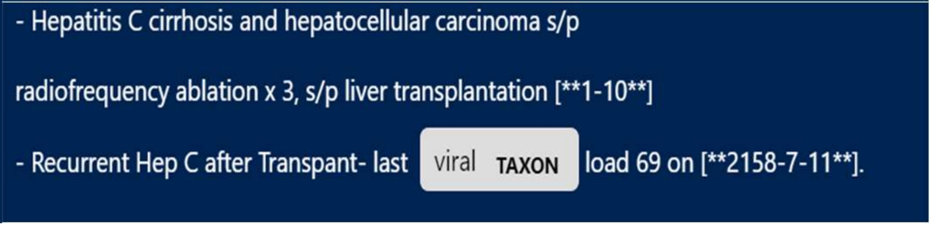
acute infarct noted on MRI [**5-15**] (left posterior frontal region SO



Congestive Heart Failure, NSTEMI, Coronary Artery Disease - s/p multiple RCA stents, Mitral Regurgitation, Diabetes Mellitus - on Insulin GGP Therapy, Hypercholesterolemia, Cerebrovascular Disease - s/p CVA, Known Carotid Disease, Right Subclavian Stenosis, Peripheral Vascular Disease, History of Humeral Fracture, GERD, Depression, Prior Bladder Surgery



(b) Persantine CHEBI MIBI on [**2183-2-4**] demonstrated a dilated left ventricle, partially reversible moderate perfusion defect in the anterior left ventricular wall, with moderate fixed deficits at the septal/inferior walls, global and left ventricular hypokinesis with an ejection fraction of 21%.

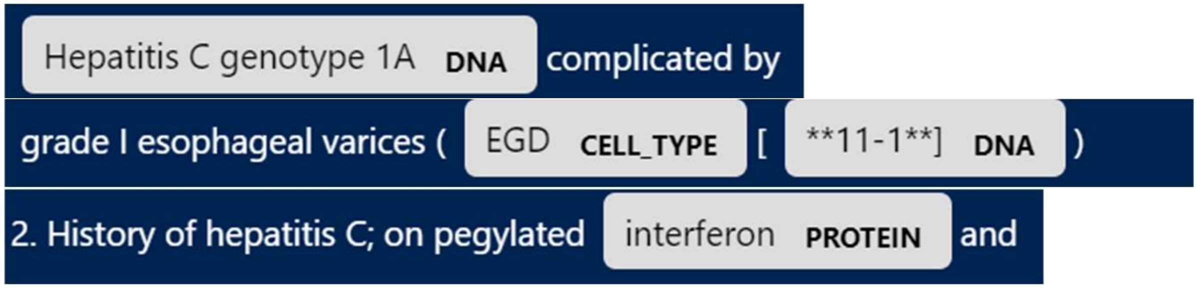


- Hepatitis C cirrhosis and hepatocellular carcinoma s/p radiofrequency ablation x 3, s/p liver transplantation [**1-10**]
- Recurrent Hep C after Transpant- last viral TAXON load 69 on [**2158-7-11**].

Scipacy: ***en_ner_craft_md***

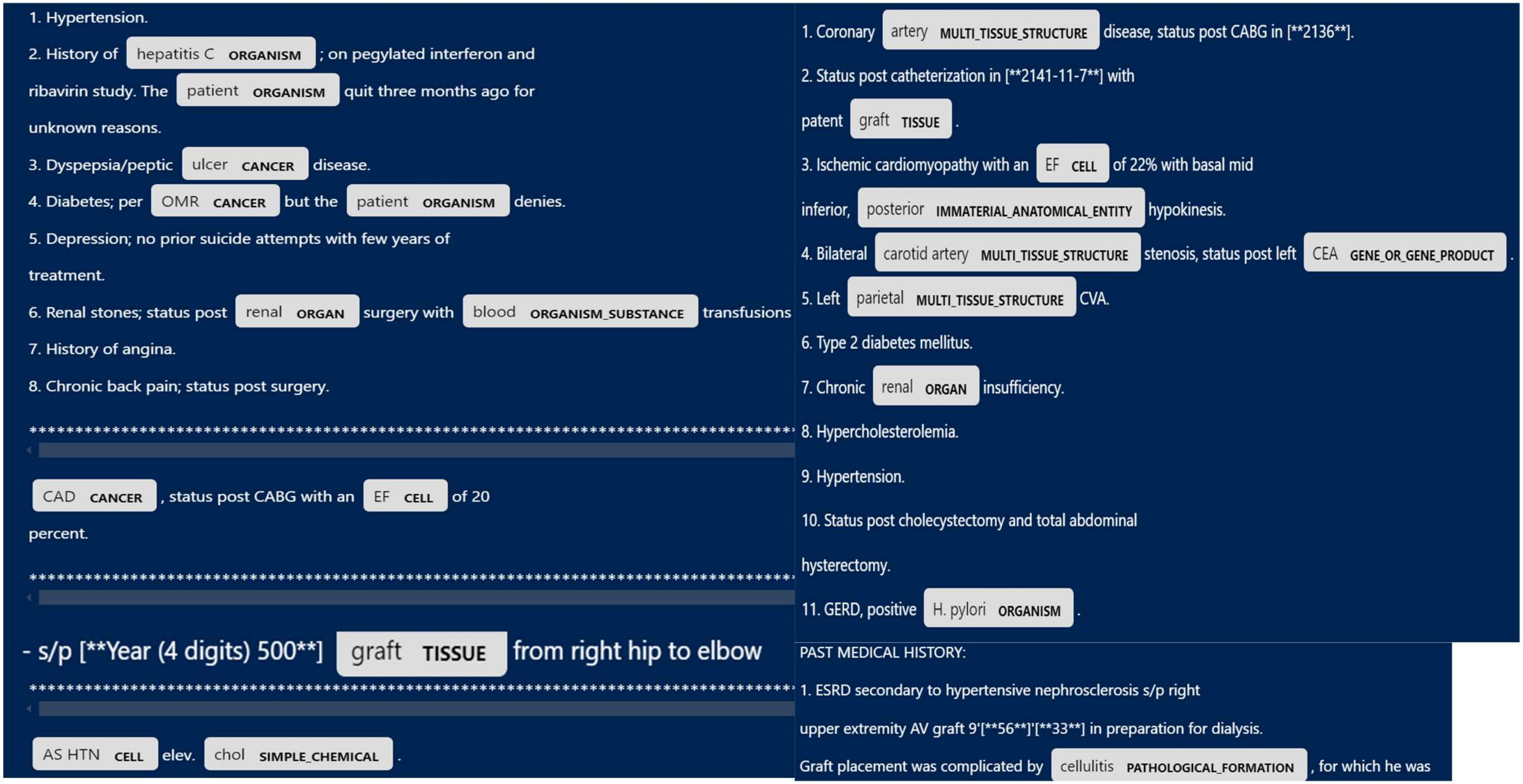
- **SO** stands for Sequence Ontology
- **CL** stands for Cell Ontology
- **CHEBI** stands for Chemical Entities of Biological Interest
- **TAXON** stands for Taxonomy
- **GCP** stands for Gene Ontology Cellular Component

Scipacy: ***en_ner_jnlpba_md***



```
Hepatitis C genotype 1A DNA complicated by
grade I esophageal varices ( EGD CELL_TYPE [ **11-1**] DNA )
2. History of hepatitis C; on pegylated interferon PROTEIN and
```

Scipacy: ***en_ner_bionlp13cg_md***



1. Hypertension.

2. History of hepatitis C ORGANISM ; on pegylated interferon and ribavirin study. The patient ORGANISM quit three months ago for unknown reasons.

3. Dyspepsia/peptic ulcer CANCER disease.

4. Diabetes; per OMR CANCER but the patient ORGANISM denies.

5. Depression; no prior suicide attempts with few years of treatment.

6. Renal stones; status post renal ORGAN surgery with blood ORGANISM_SUBSTANCE transfusions

7. History of angina.

8. Chronic back pain; status post surgery.

```
*****************************************************************************
```

CAD CANCER , status post CABG with an EF CELL of 20 percent.

```
*****************************************************************************
```

- s/p [**Year (4 digits) 500**] graft TISSUE from right hip to elbow

```
*****************************************************************************
```

AS HTN CELL elev. chol SIMPLE_CHEMICAL .

1. Coronary artery MULTI_TISSUE_STRUCTURE disease, status post CABG in [**2136**].

2. Status post catheterization in [**2141-11-7**] with patent graft TISSUE .

3. Ischemic cardiomyopathy with an EF CELL of 22% with basal mid inferior, posterior IMMATERIAL_ANATOMICAL_ENTITY hypokinesis.

4. Bilateral carotid artery MULTI_TISSUE_STRUCTURE stenosis, status post left CEA GENE_OR_GENE_PRODUCT .

5. Left parietal MULTI_TISSUE_STRUCTURE CVA.

6. Type 2 diabetes mellitus.

7. Chronic renal ORGAN insufficiency.

8. Hypercholesterolemia.

9. Hypertension.

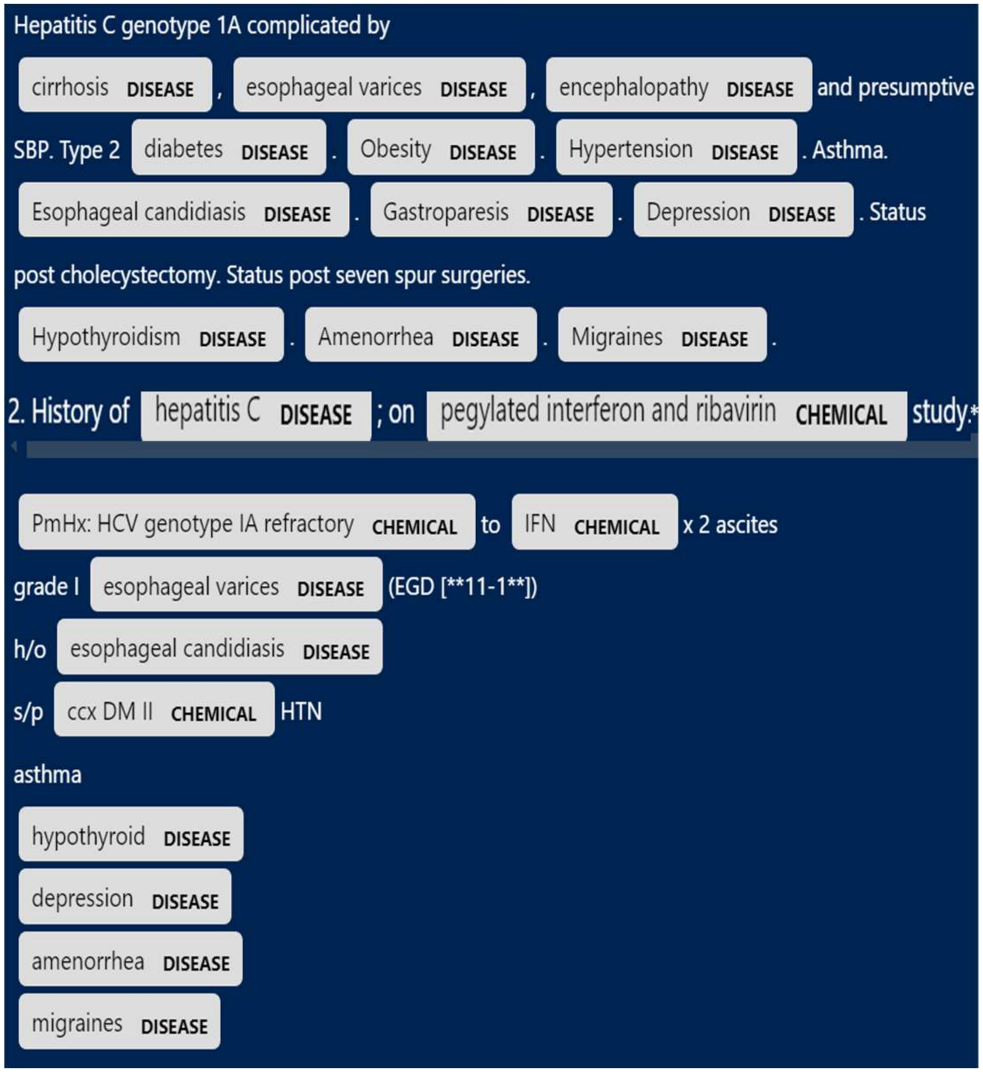
10. Status post cholecystectomy and total abdominal hysterectomy.

11. GERD, positive H. pylori ORGANISM .

PAST MEDICAL HISTORY:

1. ESRD secondary to hypertensive nephrosclerosis s/p right upper extremity AV graft 9'[**56**]'[**33**] in preparation for dialysis. Graft placement was complicated by cellulitis PATHOLOGICAL_FORMATION , for which he was

Scipacy: ***en_ner_bc5cdr_md***



Hepatitis C genotype 1A complicated by
cirrhosis DISEASE , esophageal varices DISEASE , encephalopathy DISEASE and presumptive
SBP. Type 2 diabetes DISEASE . Obesity DISEASE . Hypertension DISEASE . Asthma.
Esophageal candidiasis DISEASE . Gastroparesis DISEASE . Depression DISEASE . Status
post cholecystectomy. Status post seven spur surgeries.
Hypothyroidism DISEASE . Amenorrhea DISEASE . Migraines DISEASE .

2. History of hepatitis C DISEASE ; on pegylated interferon and ribavirin CHEMICAL study*

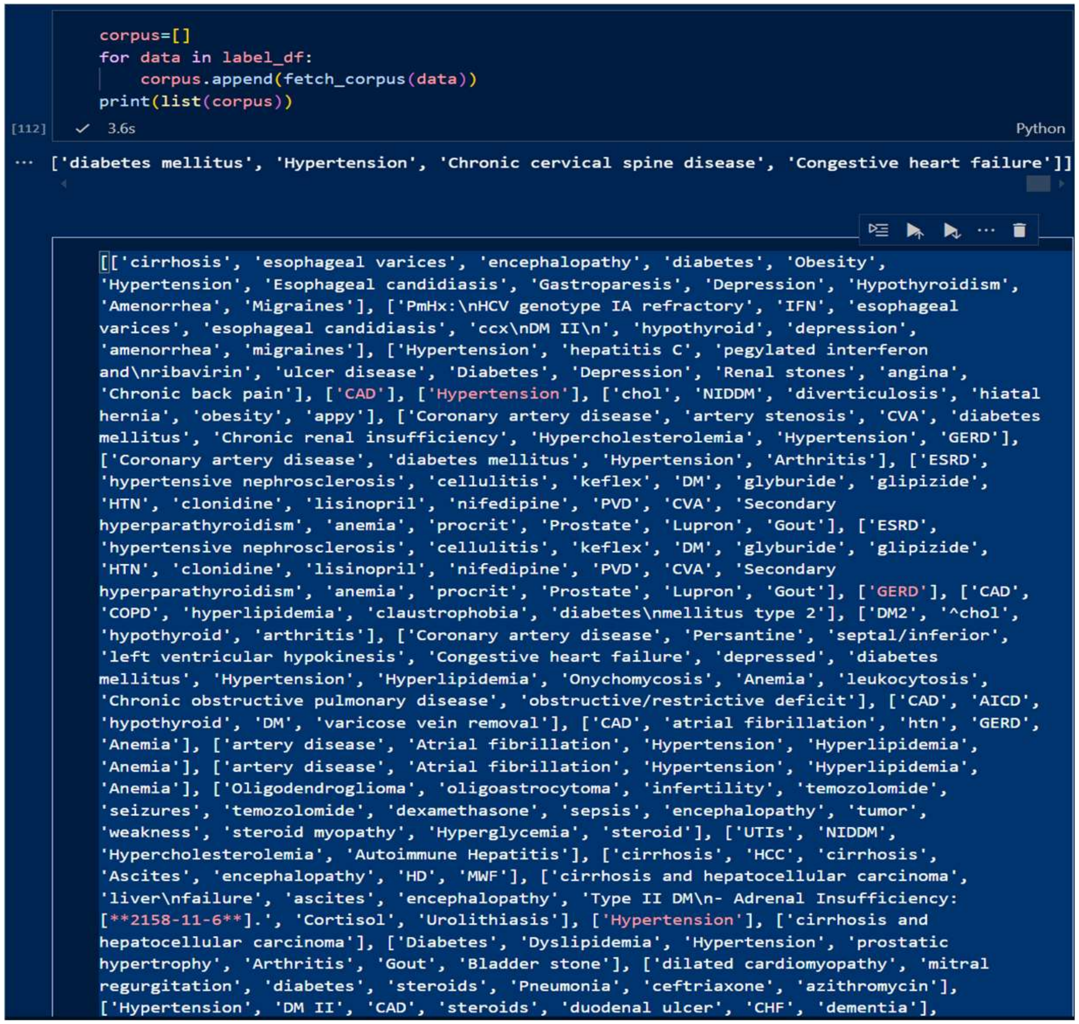
PmHx: HCV genotype IA refractory CHEMICAL to IFN CHEMICAL x 2 ascites
grade I esophageal varices DISEASE (EGD [**11-1**])
h/o esophageal candidiasis DISEASE
s/p ccx DM II CHEMICAL HTN
asthma
hypothyroid DISEASE
depression DISEASE
amenorrhea DISEASE
migraines DISEASE

Conclusion:

With the help of different spacy models we are able to extract different kinds of NERs.

Word2Vec and TSNE plot usage:

Apply **fetch_corpus()** to retrieve the list of corpus from the medical note.



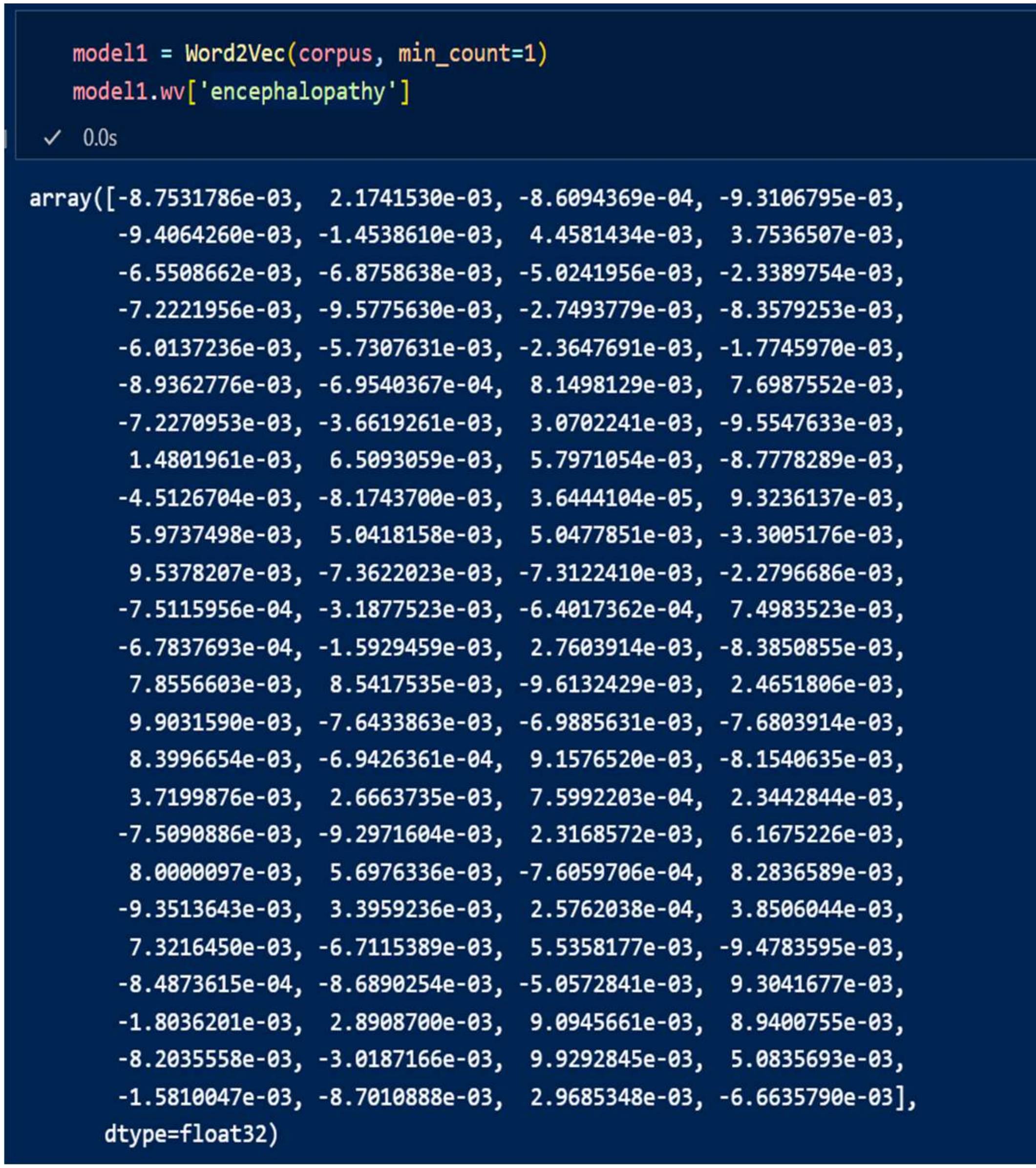
```python
corpus=[]
for data in label_df:
    corpus.append(fetch_corpus(data))
print(list(corpus))
```

```
... ['diabetes mellitus', 'Hypertension', 'Chronic cervical spine disease', 'Congestive heart failure']]
```

```
[['cirrhosis', 'esophageal varices', 'encephalopathy', 'diabetes', 'Obesity',
'Hypertension', 'Esophageal candidiasis', 'Gastroparesis', 'Depression', 'Hypothyroidism',
'Amenorrhea', 'Migraines'], ['PmHx:\nHCV genotype IA refractory', 'IFN', 'esophageal
varices', 'esophageal candidiasis', 'ccx\nDM II\n', 'hypothyroid', 'depression',
'amenorrhea', 'migraines'], ['Hypertension', 'hepatitis C', 'pegylated interferon
and\nribavirin', 'ulcer disease', 'Diabetes', 'Depression', 'Renal stones', 'angina',
'Chronic back pain'], ['CAD'], ['Hypertension'], ['chol', 'NIDDM', 'diverticulosis', 'hiatal
hernia', 'obesity', 'appy'], ['Coronary artery disease', 'artery stenosis', 'CVA', 'diabetes
mellitus', 'Chronic renal insufficiency', 'Hypercholesterolemia', 'Hypertension', 'GERD'],
['Coronary artery disease', 'diabetes mellitus', 'Hypertension', 'Arthritis'], ['ESRD',
'hypertensive nephrosclerosis', 'cellulitis', 'keflex', 'DM', 'glyburide', 'glipizide',
'HTN', 'clonidine', 'lisinopril', 'nifedipine', 'PVD', 'CVA', 'Secondary
hyperparathyroidism', 'anemia', 'procrit', 'Prostate', 'Lupron', 'Gout'], ['ESRD',
'hypertensive nephrosclerosis', 'cellulitis', 'keflex', 'DM', 'glyburide', 'glipizide',
'HTN', 'clonidine', 'lisinopril', 'nifedipine', 'PVD', 'CVA', 'Secondary
hyperparathyroidism', 'anemia', 'procrit', 'Prostate', 'Lupron', 'Gout'], ['GERD'], ['CAD',
'COPD', 'hyperlipidemia', 'claustrophobia', 'diabetes\nmellitus type 2'], ['DM2', '^chol',
'hypothyroid', 'arthritis'], ['Coronary artery disease', 'Persantine', 'septal/inferior',
'left ventricular hypokinesis', 'Congestive heart failure', 'depressed', 'diabetes
mellitus', 'Hypertension', 'Hyperlipidemia', 'Onychomycosis', 'Anemia', 'leukocytosis',
'Chronic obstructive pulmonary disease', 'obstructive/restrictive deficit'], ['CAD', 'AICD',
'hypothyroid', 'DM', 'varicose vein removal'], ['CAD', 'atrial fibrillation', 'htn', 'GERD',
'Anemia'], ['artery disease', 'Atrial fibrillation', 'Hypertension', 'Hyperlipidemia',
'Anemia'], ['artery disease', 'Atrial fibrillation', 'Hypertension', 'Hyperlipidemia',
'Anemia'], ['Oligodendroglioma', 'oligoastrocytoma', 'infertility', 'temozolomide',
'seizures', 'temozolomide', 'dexamethasone', 'sepsis', 'encephalopathy', 'tumor',
'weakness', 'steroid myopathy', 'Hyperglycemia', 'steroid'], ['UTIs', 'NIDDM',
'Hypercholesterolemia', 'Autoimmune Hepatitis'], ['cirrhosis', 'HCC', 'cirrhosis',
'Ascites', 'encephalopathy', 'HD', 'MWF'], ['cirrhosis and hepatocellular carcinoma',
'liver\nfailure', 'ascites', 'encephalopathy', 'Type II DM\n- Adrenal Insufficiency:
[**2158-11-6**].', 'Cortisol', 'Urolithiasis'], ['Hypertension'], ['cirrhosis and
hepatocellular carcinoma'], ['Diabetes', 'Dyslipidemia', 'Hypertension', 'prostatic
hypertrophy', 'Arthritis', 'Gout', 'Bladder stone'], ['dilated cardiomyopathy', 'mitral
regurgitation', 'diabetes', 'steroids', 'Pneumonia', 'ceftriaxone', 'azithromycin'],
['Hypertension', 'DM II', 'CAD', 'steroids', 'duodenal ulcer', 'CHF', 'dementia'],
```

Code Snippet:

The resultant presents the similarity logits for the word ‘encephalopathy’

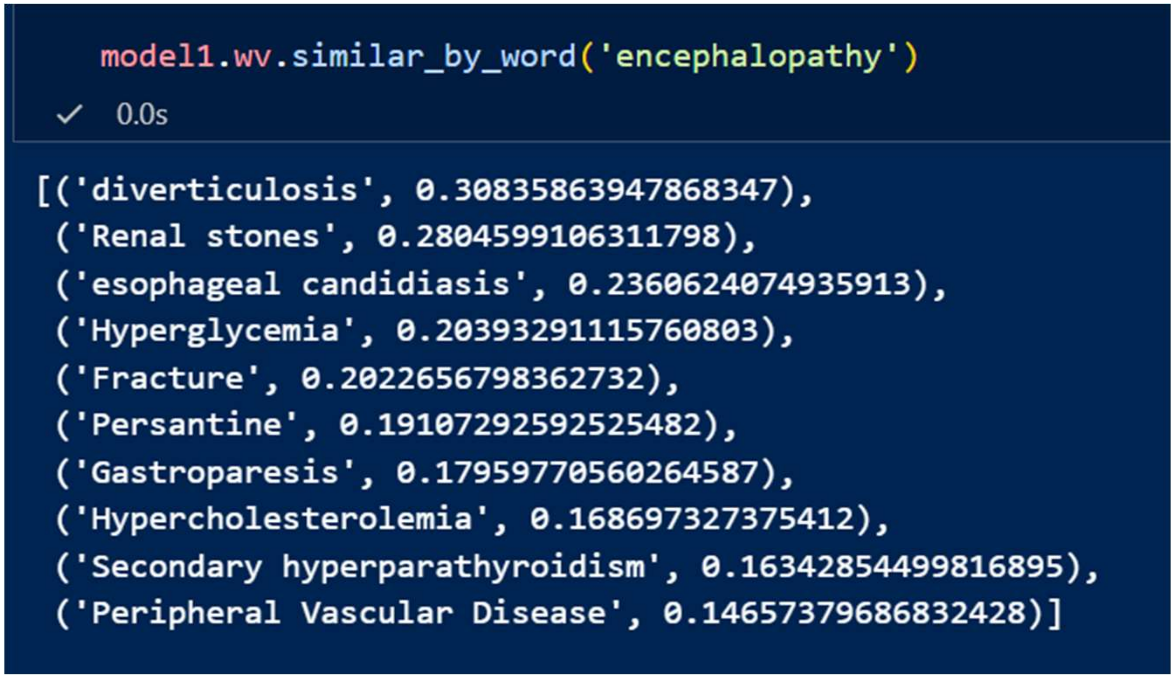


```
model1 = Word2Vec(corpus, min_count=1)
model1.wv['encephalopathy']
```

✓ 0.0s

```
array([-8.7531786e-03,  2.1741530e-03, -8.6094369e-04, -9.3106795e-03,
       -9.4064260e-03, -1.4538610e-03,  4.4581434e-03,  3.7536507e-03,
       -6.5508662e-03, -6.8758638e-03, -5.0241956e-03, -2.3389754e-03,
       -7.2221956e-03, -9.5775630e-03, -2.7493779e-03, -8.3579253e-03,
       -6.0137236e-03, -5.7307631e-03, -2.3647691e-03, -1.7745970e-03,
       -8.9362776e-03, -6.9540367e-04,  8.1498129e-03,  7.6987552e-03,
       -7.2270953e-03, -3.6619261e-03,  3.0702241e-03, -9.5547633e-03,
        1.4801961e-03,  6.5093059e-03,  5.7971054e-03, -8.7778289e-03,
       -4.5126704e-03, -8.1743700e-03,  3.6444104e-05,  9.3236137e-03,
        5.9737498e-03,  5.0418158e-03,  5.0477851e-03, -3.3005176e-03,
        9.5378207e-03, -7.3622023e-03, -7.3122410e-03, -2.2796686e-03,
       -7.5115956e-04, -3.1877523e-03, -6.4017362e-04,  7.4983523e-03,
       -6.7837693e-04, -1.5929459e-03,  2.7603914e-03, -8.3850855e-03,
        7.8556603e-03,  8.5417535e-03, -9.6132429e-03,  2.4651806e-03,
        9.9031590e-03, -7.6433863e-03, -6.9885631e-03, -7.6803914e-03,
        8.3996654e-03, -6.9426361e-04,  9.1576520e-03, -8.1540635e-03,
        3.7199876e-03,  2.6663735e-03,  7.5992203e-04,  2.3442844e-03,
       -7.5090886e-03, -9.2971604e-03,  2.3168572e-03,  6.1675226e-03,
        8.0000097e-03,  5.6976336e-03, -7.6059706e-04,  8.2836589e-03,
       -9.3513643e-03,  3.3959236e-03,  2.5762038e-04,  3.8506044e-03,
        7.3216450e-03, -6.7115389e-03,  5.5358177e-03, -9.4783595e-03,
       -8.4873615e-04, -8.6890254e-03, -5.0572841e-03,  9.3041677e-03,
       -1.8036201e-03,  2.8908700e-03,  9.0945661e-03,  8.9400755e-03,
       -8.2035558e-03, -3.0187166e-03,  9.9292845e-03,  5.0835693e-03,
       -1.5810047e-03, -8.7010888e-03,  2.9685348e-03, -6.6635790e-03],
      dtype=float32)
```

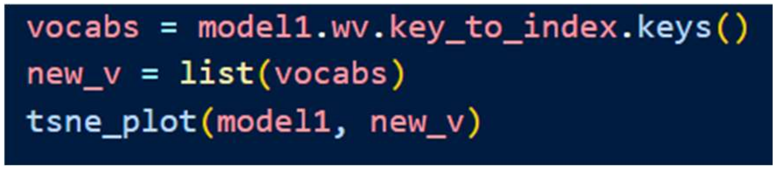
The code snippet shows the similar words that the model contains within it.



```
model1.wv.similar_by_word('encephalopathy')
```

```
[('diverticulosis', 0.30835863947868347),
 ('Renal stones', 0.2804599106311798),
 ('esophageal candidiasis', 0.2360624074935913),
 ('Hyperglycemia', 0.20393291115760803),
 ('Fracture', 0.2022656798362732),
 ('Persantine', 0.19107292592525482),
 ('Gastroparesis', 0.17959770560264587),
 ('Hypercholesterolemia', 0.168697327375412),
 ('Secondary hyperparathyroidism', 0.16342854499816895),
 ('Peripheral Vascular Disease', 0.14657379686832428)]
```

TSNE plot:

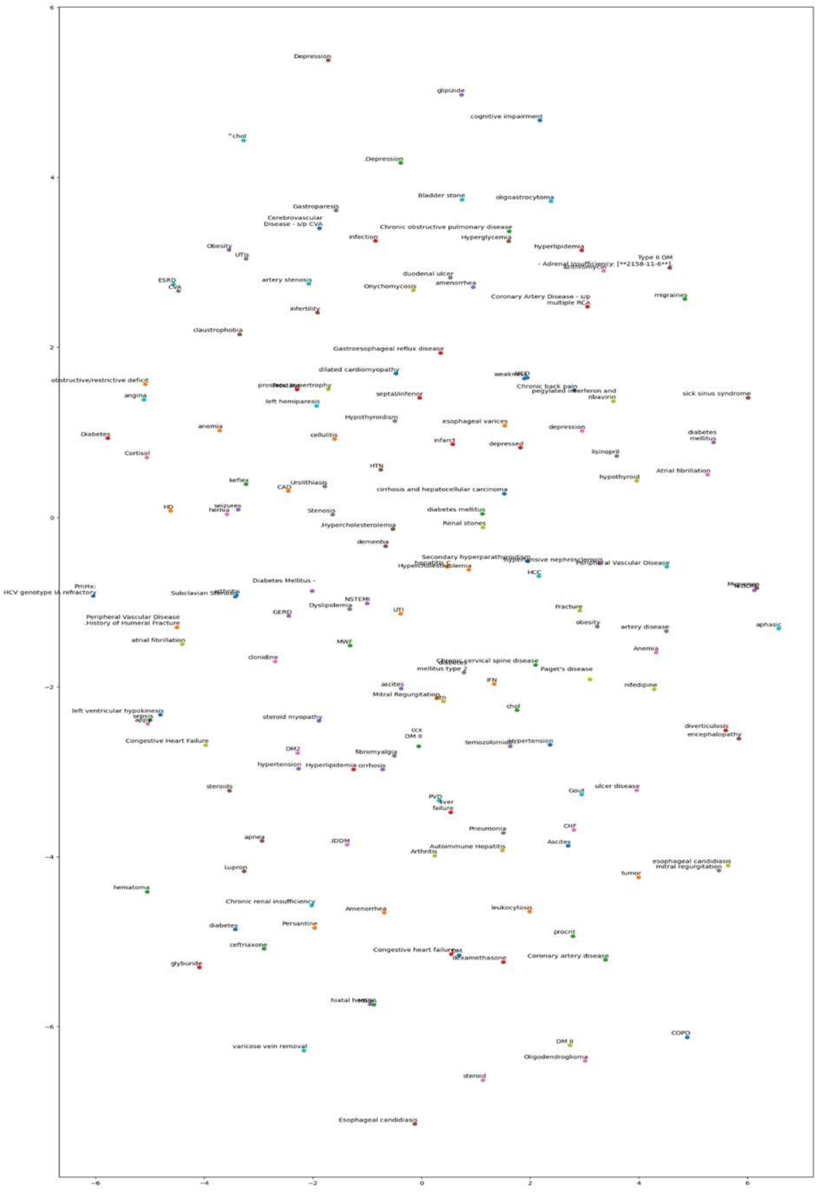


```
vocabs = model1.wv.key_to_index.keys()
new_v = list(vocabs)
tsne_plot(model1, new_v)
```

The corpus are plotted in tsne word embedding as per their logits.

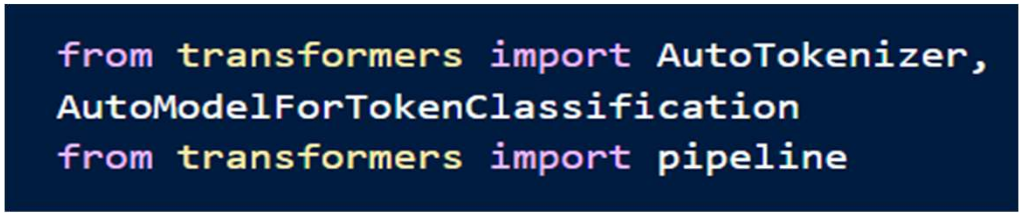
SciSpacy TSNE plot:

https://github.com/dalwari/mimic-iii-clinical-database-demo-1.4/blob/main/tsne_plot_diag.png


6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6
Depression
glipizide
cognitive impairment
^chol
.Depression
Bladder stone
oligoastrocytoma
Gastroparesis
Cerebrovascular
Disease - s/p CVA
Chronic obstructive pulmonary disease
infection
Hyperglycemia
Obesity
UTIs
hyperlipidemia
Type II DM
- Adrenal insufficiency: [**2158-11-6**]
azithromycin
duodenal ulcer
ESRD
CVA
artery stenosis
Onychomycosis
amenorrhea
Coronary Artery Disease - s/p
multiple RCA
migraines
infertility
claustrophobia
Gastroesophageal reflux disease
dilated cardiomyopathy
weakness
obstructive/restrictive deficit
prostatic hypertrophy
Tobacco
Chronic back pain
pegylated interferon and
ribavirin
sick sinus syndrome
angina
septal/inferior
left hemiparesis
Hypothyroidism
esophageal varices
depression
anemia
diabetes
mellitus
Diabetes
cellulitis
infarct
depressed
Cortisol
lisinopril
HTN
hypothyroid
Atrial fibrillation
keflex
Urolithiasis
CAD
cirrhosis and hepatocellular carcinoma
HD
seizures
hernia
Stenosis
diabetes mellitus
Renal stones
.Hypercholesterolemia
dementia
Secondary hyperparathyroidism
hepatitis C
Hypercholesterolemia
Hypertensive nephrosclerosis
Peripheral Vascular Disease
HCC
Migraines
PmHx:
HCV genotype IA refractory
Diabetes Mellitus -
arthritis
Subclavian Stenosis
NSTEMI
Dyslipidemia
GERD
UTI
Fracture
Peripheral Vascular Disease
.History of Humeral Fracture
obesity
artery disease
aphasic
atrial fibrillation
MWF
Anemia
clonidine
Chronic cervical spine disease
diabetes
mellitus type 2
Paget's disease
IFN
ascites
nifedipine
Mitral Regurgitation
htn
chol
left ventricular hypokinesis
sepsis
apap
steroid myopathy
diverticulosis
encephalopathy
ccx
DM II
Congestive Heart Failure
temozolomide
Hypertension
DM2
fibromyalgia
hypertension
Hyperlipidemia
cirrhosis
steroids
ulcer disease
Gout
PVD
liver
failure
Pneumonia
CHF
apnea
.IDDM
Autoimmune Hepatitis
Arthritis
Ascites
esophageal candidiasis
mitral regurgitation
Lupron
tumor
hematoma
Chronic renal insufficiency
Amenorrhea
leukocytosis
diabetes
Persantine
procrit
ceftriaxone
Congestive heart failure
DM
dexamethasone
Coronary artery disease
glyburide
hiatal hernia
MSSA
COPD
varicose vein removal
DM II
Oligodendroglioma
steroid
Esophageal candidiasis

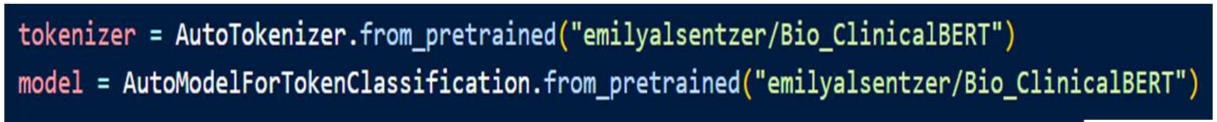
Clinical BERT:

- Import transformers and classification package.

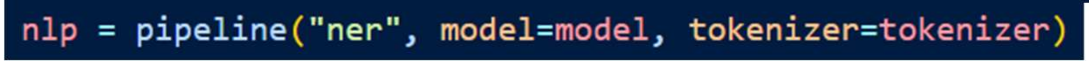


```
from transformers import AutoTokenizer,
AutoModelForTokenClassification
from transformers import pipeline
```

- Initialize model.



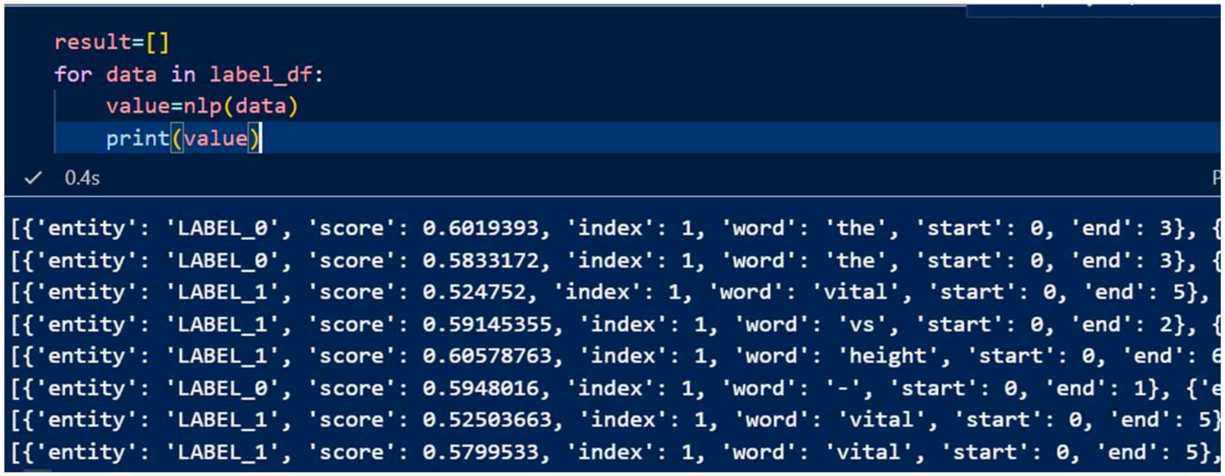
```
tokenizer = AutoTokenizer.from_pretrained("emilyalsentzer/Bio_ClinicalBERT")
model = AutoModelForTokenClassification.from_pretrained("emilyalsentzer/Bio_ClinicalBERT")
```



```
nlp = pipeline("ner", model=model, tokenizer=tokenizer)
```

- Load dataset and examine ‘PHYSICAL EXAMINATION’ sub passage.

- The output displays label and score for each word in a sentence.
- The label is categorized by ClinicalBERT into LABEL_0 and LABEL_1
- The score is given to each word in that label category.

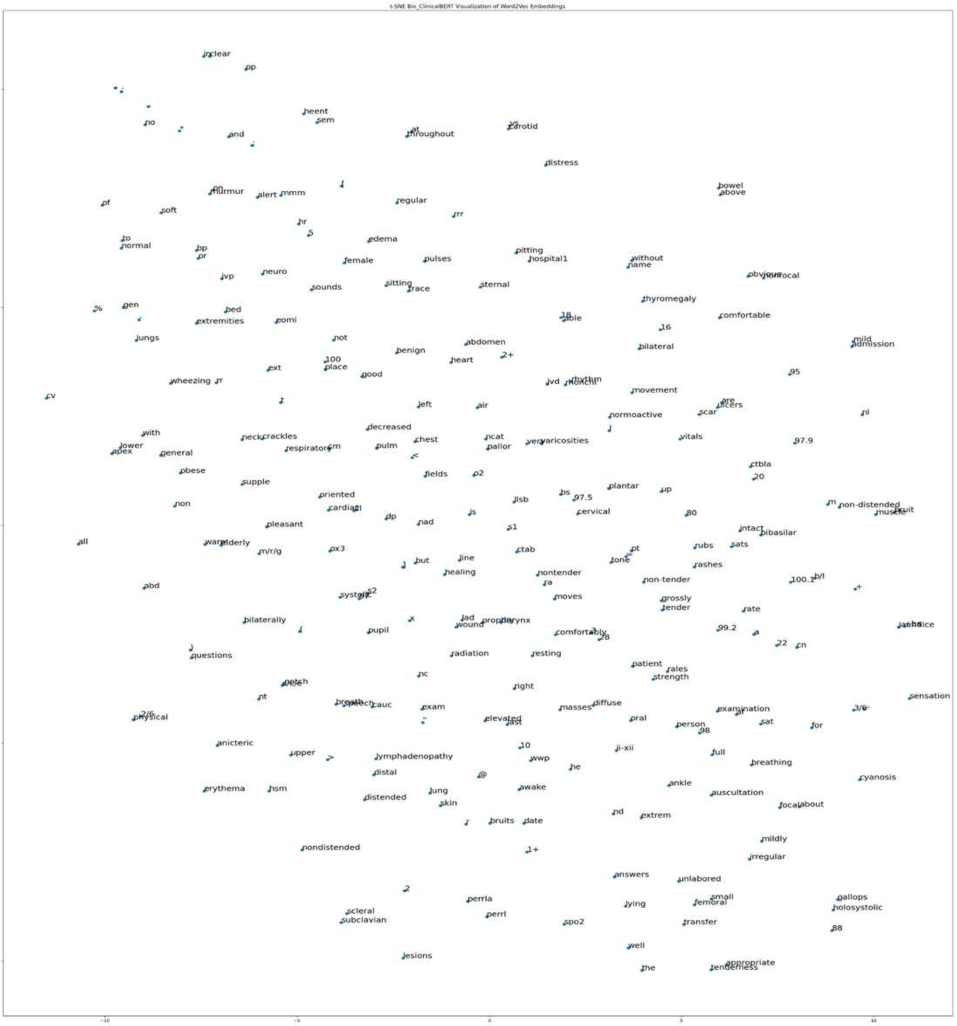


```
result=[]
for data in label_df:
    value=nlp(data)
    print(value)
```

```
[{'entity': 'LABEL_0', 'score': 0.6019393, 'index': 1, 'word': 'the', 'start': 0, 'end': 3}, {
[{'entity': 'LABEL_0', 'score': 0.5833172, 'index': 1, 'word': 'the', 'start': 0, 'end': 3}, {
[{'entity': 'LABEL_1', 'score': 0.524752, 'index': 1, 'word': 'vital', 'start': 0, 'end': 5},
[{'entity': 'LABEL_1', 'score': 0.59145355, 'index': 1, 'word': 'vs', 'start': 0, 'end': 2}, {
[{'entity': 'LABEL_1', 'score': 0.60578763, 'index': 1, 'word': 'height', 'start': 0, 'end': 6
[{'entity': 'LABEL_0', 'score': 0.5948016, 'index': 1, 'word': '-', 'start': 0, 'end': 1}, {'e
[{'entity': 'LABEL_1', 'score': 0.52503663, 'index': 1, 'word': 'vital', 'start': 0, 'end': 5}
[{'entity': 'LABEL_1', 'score': 0.5799533, 'index': 1, 'word': 'vital', 'start': 0, 'end': 5},
```

ClinicalBERT TSNE visualization: For Medication in medical notes

https://github.com/dalwari/mimic-iii-clinical-database-demo-1.4/blob/main/clinical_bert_tsne_plot.png


t-SNE Bio_ClinicalBERT Visualization of Word2Vec Embeddings
unclear
op
heent
sem
no
and
at
throughout
vs
carotid
distress
murmur
alert
mmm
regular
bowel
above
of
soft
rrr
hr
to
normal
bp
pr
edema
female
pulses
pitting
hospital1
without
name
jvp
neuro
sounds
sitting
trace
sternal
obvious
nonfocal
thyromegaly
gen
bed
extremities
eomi
able
comfortable
lungs
not
benign
abdomen
heart
bilateral
mild
admission
place
ext
good
wheezing
jvd
rhythm
movement
cv
left
air
are
ulcers
scar
normoactive
nl
with
lower
apex
general
neck
crackles
respiratory
cm
decreased
chest
pulm
ncat
pallor
very
varicosities
vitals
obese
supple
fields
o2
ctbla
oriented
cardiac
plantar
up
non
pleasant
dp
nad
js
llsb
bs
cervical
m
non-distended
muscle
bruit
s1
intact
bibasilar
all
warm
elderly
m/r/g
ox3
but
line
healing
ctab
pt
tone
rubs
sats
rashes
nontender
non-tender
b/l
abd
s2
moves
grossly
tender
rate
bilaterally
pupil
x
lad
wound
larynx
comfortably
a
cn
questions
radiation
resting
patient
rales
strength
right
sensation
nt
breath
speech
cauc
exam
masses
diffuse
oral
examination
sat
for
3/6
2/6
physical
elevated
last
person
anicteric
ii-xii
full
breathing
upper
lymphadenopathy
wwp
he
distal
@
cyanosis
erythema
hsm
lung
skin
awake
ankle
auscultation
focal
about
distended
nd
extrem
bruits
date
mildly
irregular
nondistended
answers
unlabored
perrla
lying
small
femoral
gallops
holosystolic
scleral
subclavian
perrl
spo2
transfer
well
lesions
the
tenderness
appropriate