Idrise Abdi (ITA259)

Analysis of MIMIC HFpEF Data with Spacy

1. Extract Text from HF Diagnoses

- MIMIC Notes and Diagnoses data are loaded (csv – GoogleDrive)
- Merge DataFrames (DF) on Admission ID (HADM_ID)
- Filter the merged DF for AC Diastolic Heart Failure (Preserved Ejection)
  - HFpEF diagnosis was selected in part because of its lower corpus size and because it is in the HF disease-domain on which my assignments have been focused.
- Spacy is loaded with the medium model
- Text is extracted from the DF by selecting the first 1000k rows and selecting the ‘TEXT’ field.
- For each MIMIC “Note Entry”, a Spacy doc is created and the entities are extracted to a list

Entity Extraction

- Each Spacy doc has the . Ents property/tuple (text,label)
- These are listed and grouped with counts per group

HFpEF Extracted Entities – en_ner_bc5cdr_md

entity_med_grouped	Count
Pt	1215
pain	412
CHF	339
lasix	326
NO	325
pneumonia	267
edema	263
heparin	231
cough	217
SOB	210
Lasix	202
pulmonary edema	199

CAD	181
pleural effusions	177
COPD	169
pleural effusion	160
atelectasis	159
NSR	144
respiratory distress	140
chest pain	136
HTN	134
pneumothorax	132
w/	130
O2	130
Heparin	125
hypertension	122

mitral regurgitation	122
DVT	122
NG	119
creatinine	118
Allergies	118
fentanyl	116
LLL	115
shortness of breath	114
bleeding	110
K	108
oxygen	106
pericardial effusion	104
hypotension	103
steroids	99
fever	99
effusion	99

HFpEF Extracted Entities with Large Sci–

en_core_sci_lg

entity	type
Pt	1372
patient	981
day	569
increased	534
PT	503
PO	500
BP	494
HR	489
Reason	487
PM	456
Tablet	401
Patient	398
AM	381
REASON	372
CT	357
Plan	337
BS	336
CHF	328
CV	317
O2	299
unchanged	289
stable	286
RR	282
NO	270
evidence	269

HFpEF Extracted Entities with bio-bert

en_biobert_ner_symptom

SOB	46
cough	31
pain	27
shortness of breath	18
CP	13
sob	13
chest pain	8
fever	8
PAIN	7
seizure activity	7
SHORTNESS OF BREATH	7
swelling	7

Word2Vec embeddings

- Word2Vec is loaded with a corpus loaded from tokens from the mimic hf text dataframe
- Word similarities are calculated on some sample data using similar_by_word
- The tsne_plot function from the canvas page was reused to plot the entities embeddings and entity labels

Word2Vec Embeddings

More relevant results with df_nlp_sci_large

```
[64] model1.wv.similar_by_word('Sinus rhythm')

[('NO', 0.9896019697189331),
 ('stool', 0.9895113110542297),
 ('CV', 0.9892734885215759),
 ('HR', 0.9889732599258423),
 ('ID', 0.988877534866333),
 ('goal', 0.9887533187866211),
 ('CHF', 0.9887236952781677),
 ('patient', 0.988703191280365),
 ('day', 0.9886724352836609),
 ('diuresis', 0.9886477589607239)]
```

```
[65] modellarge.wv.similar_by_word('Sinus rhythm')

[('ST-T wave abnormalities', 0.9995008111000061),
 ('anterior', 0.999462902545929),
 ('ischemia', 0.999433696269989),
 ('consulted', 0.9994246363639832),
 ('results', 0.9993900656700134),
 ('Hypertension', 0.9993869662284851),
 ('HCO3', 0.9993867874145508),
 ('removed', 0.999347448348999),
 ('resolved', 0.9993336796760559),
 ('peripheral', 0.9993318915367126)]
```

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

[('changes', 0.9718748331069946),
 ('night', 0.9711485505104065),
 ('stool', 0.9708017706871033),
 ('BP', 0.9707691073417664),
 ('RR', 0.970663845539093),
 ('tolerated', 0.9704496264457703),
 ('CVP', 0.970432698726654),
 ('moderate', 0.9702677130699158),
 ('NO', 0.9702579975128174),
 ('lasix', 0.9700937867164612)]
```

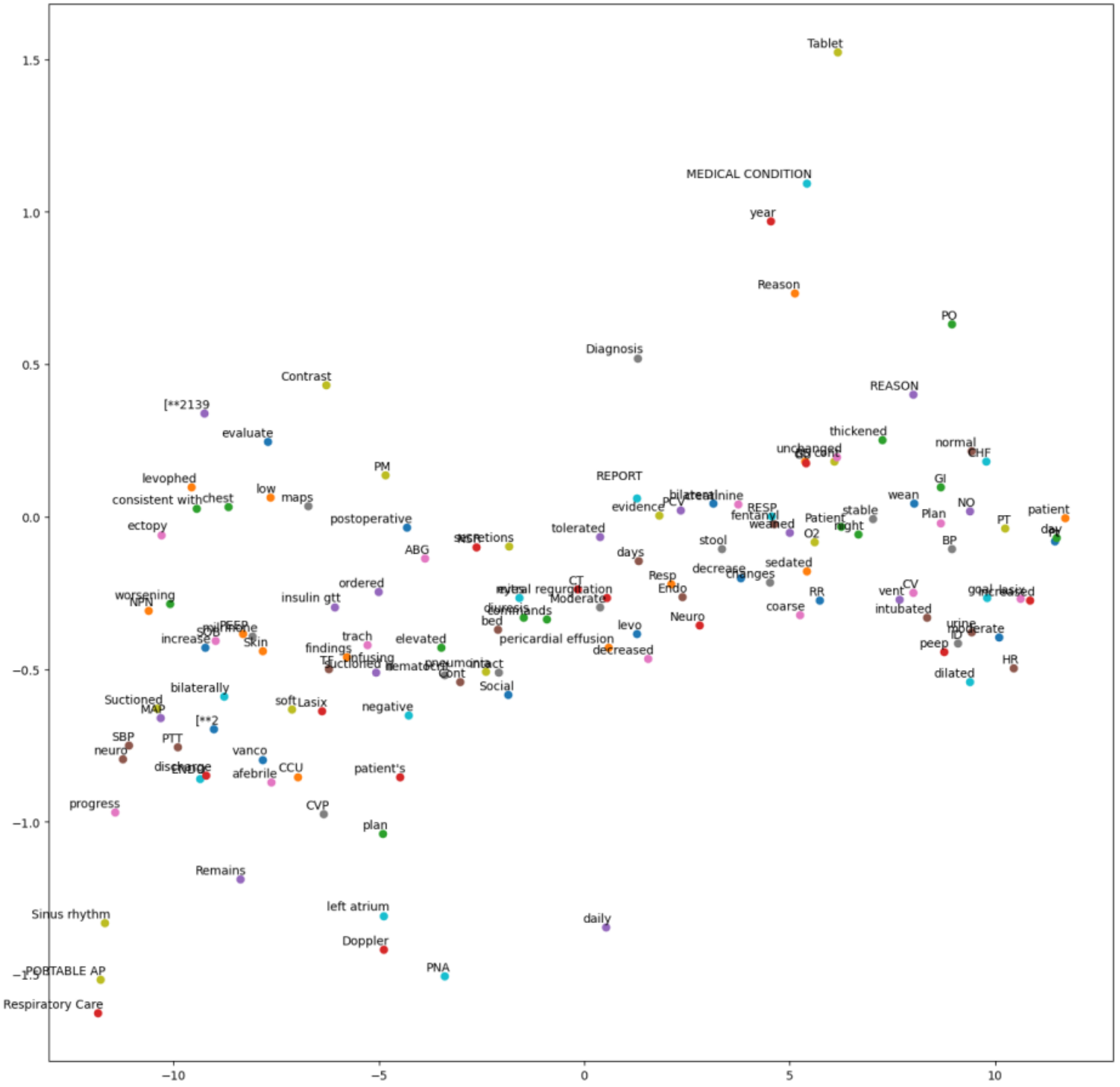
```
6] modellarge.wv.similar_by_word('ischemia')

[('Hypertension', 0.9995061755180359),
 ('consulted', 0.9994360208511353),
 ('Sinus rhythm', 0.9994336366653442),
 ('clinically', 0.9994090795516968),
 ('anterior', 0.9993607401847839),
 ('died', 0.9992967247962952),
 ('treated', 0.9992964267730713),
 ('symptoms', 0.9992923736572266),
 ('ST-T wave abnormalities', 0.9992920756340027),
 ('resolved', 0.9992830753326416)]
```

Word2Vec TSNE Plot

— en_ner_bc5cdr_md

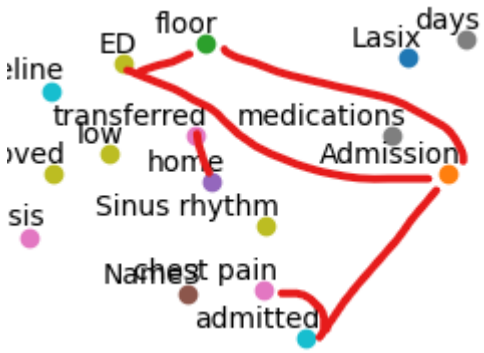
Not so relevant clusters



Word2Vec TSNE Plot

— df_nlp_sci_large

More relevant results with df_nlp_sci_large



Word2Vec TSNE Plot

—

en_biobert_ner_symptom

Decent results with en_biobert_ner_symptom

T-SNE with BioBert