



# MIMIC NLP

SUJAY GOPINATHAN (SG59258)

# User Story

- ▶ Find the most common disease from the MIMIC-III database and do NLP on the notes with the ICD9 code corresponding to the disease.
- ▶ Will use DIAGNOSES\_ICD, NOTEEVENTS, D\_ICD\_DIAGNOSES from the MIMIC-III database <https://physionet.org/content/mimiciii/1.4/>
- ▶ Use Spacy, SciSpacy, Word2Vec, tSNE plots and medSpacy as NLP tools.
- ▶ Github URL: <https://github.com/sujaycloud/aih>

# Read the MIMIC-III files

1. Download the DIAGNOSES\_ICD, NOTEEVENTS and D\_ICD\_DIAGNOSES csv files from the MIMIC-III database.
2. Read the csv files using pandas
3. Find the most common disease identified by the ICD9\_CODE from the DIAGNOSES\_ICD file.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 651047 entries, 0 to 651046
Data columns (total 5 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   ROW_ID      651047 non-null  int64
1   SUBJECT_ID  651047 non-null  int64
2   HADM_ID     651047 non-null  int64
3   SEQ_NUM     651000 non-null  float64
4   ICD9_CODE   651000 non-null  object
dtypes: float64(1), int64(3), object(1)
memory usage: 24.8+ MB
['4019']
```

4. ICD9\_CODE = 4019 is the most common disease
5. Create a new csv (disease\_notes.csv) for all notes of patients with ICD9 code as 4019.

# Spacy Assignment

1. Read the disease\_notes.csv file, preprocess the data and then load into the model using Spacy.
2. Print out the lines from each note

```
title: protected section demographics day of intubation: day of mechanical ventilation: ideal body weight: none ideal tidal volume: 2.2 ml/kg airway airway placement data known difficult intubation: no tube
*****
intestine, perforation of (perforation of hollow viscus) assessment: action: response: plan: protected section nsg note error. protected section error entered by: , rn on: :
*****
title: protected section demographics day of intubation: day of mechanical ventilation: ideal body weight: none ideal tidal volume: 2.2 ml/kg airway airway placement data known difficult intubation: no tube
*****
intestine, perforation of (perforation of hollow viscus) assessment: action: response: plan: protected section nsg note error. protected section error entered by: , rn on: :
*****
```

3. Print out the tokens and NER's

```
title NOUN
: PUNCT
protected VERB
section NOUN
demographics NOUN
day NOUN
of ADP
intubation NOUN
: PUNCT
day NOUN
of ADP
mechanical ADJ
ventilation NOUN
: PUNCT
ideal ADJ
body NOUN
weight NOUN
: PUNCT
none NOUN
```

```
mm lung 275 282 PERSON
next - hours 792 804 TIME
*****
mm lung 275 282 PERSON
next - hours 792 804 TIME
*****
hour 17 21 TIME
midazolam 336 345 PERSON
hours 638 643 TIME
cmho plateau 1105 1117 PERSON
cmho spo 1119 1127 PERSON
hct plt 1568 1575 PERSON
mg++ 1837 1841 PERSON
# 1949 1950 CARDINAL
mrsa pna 2494 2502 ORG
this afternoon 2532 2546 TIME
```

# Spacy Entity Visualization

title: protected section demographics day of intubation: day of mechanical ventilation: ideal body weight: none ideal tidal volume: // ml/kg airway airway placement data known difficult intubation: no tube type tracheostomy tube: type: perc trach manufacturer: portex size: mm lung PERSON sounds rll lung sounds: diminished rul lung sounds: rhonchi lul lung sounds: rhonchi ill lung sounds: diminished secretions sputum color / consistency: blood tinged / thick sputum source/amount: suctioned / moderate ventilation assessment level of breathing assistance: continuous invasive ventilation. visual assessment of breathing pattern: normal quiet breathing. assessment of breathing comfort: no claim of dyspnea. invasive ventilation assessment: trigger work assessment: triggering synchronously plan next - hours TIME : patient received on a/c. weaned to psv / through the shift. tolerating well. plan to continue on psv as tolerated at this time. reason for continuing current ventilatory support: underlying illness not resolved. protected section error entered by: , rt on: :

\*\*\*\*\*

intestine, perforation of (perforation of hollow viscus) assessment: action: response: plan: protected section nsg note error. protected section error entered by: , rn on: :

\*\*\*\*\*

title: protected section demographics day of intubation: day of mechanical ventilation: ideal body weight: none ideal tidal volume: // ml/kg airway airway placement data known difficult intubation: no tube type tracheostomy tube: type: perc trach manufacturer: portex size: mm lung PERSON sounds rll lung sounds: diminished rul lung sounds: rhonchi lul lung sounds: rhonchi ill lung sounds: diminished secretions sputum color / consistency: blood tinged / thick sputum source/amount: suctioned / moderate ventilation assessment level of breathing assistance: continuous invasive ventilation. visual assessment of breathing pattern: normal quiet breathing. assessment of breathing comfort: no claim of dyspnea. invasive ventilation assessment: trigger work assessment: triggering synchronously plan next - hours TIME : patient received on a/c. weaned to psv / through the shift. tolerating well. plan to continue on psv as tolerated at this time. reason for continuing current ventilatory support: underlying illness not resolved. protected section error entered by: , rt on: :

# SciSpacy Assignment

1. Do the same exercise as Spacy with an untrained model (en\_core\_sci\_md) and visualize the named entities

title: protected section ENTITY demographics ENTITY day ENTITY of intubation ENTITY : day ENTITY of mechanical ventilation ENTITY : ideal body weight ENTITY : none ideal tidal volume ENTITY : // ml/kg ENTITY airway ENTITY airway ENTITY placement ENTITY data ENTITY known difficult intubation ENTITY : no tube type ENTITY tracheostomy tube ENTITY : type: perc trach manufacturer ENTITY : portex size ENTITY : mm lung sounds ENTITY rll lung sounds: diminished ENTITY rul lung sounds ENTITY : rhonchi lul lung sounds ENTITY : rhonchi llr lung sounds ENTITY : diminished secretions ENTITY sputum color ENTITY / consistency ENTITY : blood tinged ENTITY / thick ENTITY sputum source/amount ENTITY : suctioned ENTITY / moderate ENTITY ventilation assessment ENTITY level ENTITY of breathing assistance ENTITY : continuous invasive ventilation ENTITY . visual assessment ENTITY of breathing pattern ENTITY : normal quiet breathing ENTITY . assessment ENTITY of breathing comfort ENTITY : no claim of dyspnea ENTITY . invasive ENTITY ventilation assessment: trigger work assessment ENTITY : triggering synchronously plan ENTITY next ENTITY - hours ENTITY : patient ENTITY received on a/c ENTITY . weaned ENTITY to psv ENTITY / through the shift ENTITY . tolerating well. plan ENTITY to continue on psv ENTITY as tolerated ENTITY at this time ENTITY . reason for continuing current ENTITY ventilatory ENTITY support: underlying illness ENTITY not resolved. protected section error ENTITY entered by: , rt on: :

2. SciSpacy is able to do much better than Spacy
3. Use NER model (en\_ner\_bc5cdr\_md) to identify diseases

title: protected section demographics day of intubation: day of mechanical ventilation: ideal body weight: none ideal tidal volume: // ml/kg airway airway placement data known difficult intubation: no tube type tracheostomy tube: type: perc trach manufacturer: portex size: mm lung sounds rll lung sounds: diminished rul lung sounds: rhonchi lul lung sounds: rhonchi llr lung sounds: diminished secretions sputum color / consistency: blood tinged / thick sputum source/amount: suctioned / moderate ventilation assessment level of breathing assistance: continuous invasive ventilation. visual assessment of breathing pattern: normal quiet breathing. assessment of breathing comfort: no claim of dyspnea DISEASE . invasive ventilation assessment: trigger work assessment: triggering synchronously plan next - hours: patient received on a/c. weaned to psv / through the shift. tolerating well. plan to continue on psv as tolerated at this time. reason for continuing current ventilatory support: underlying illness not resolved. protected section error entered by: , rt on: :

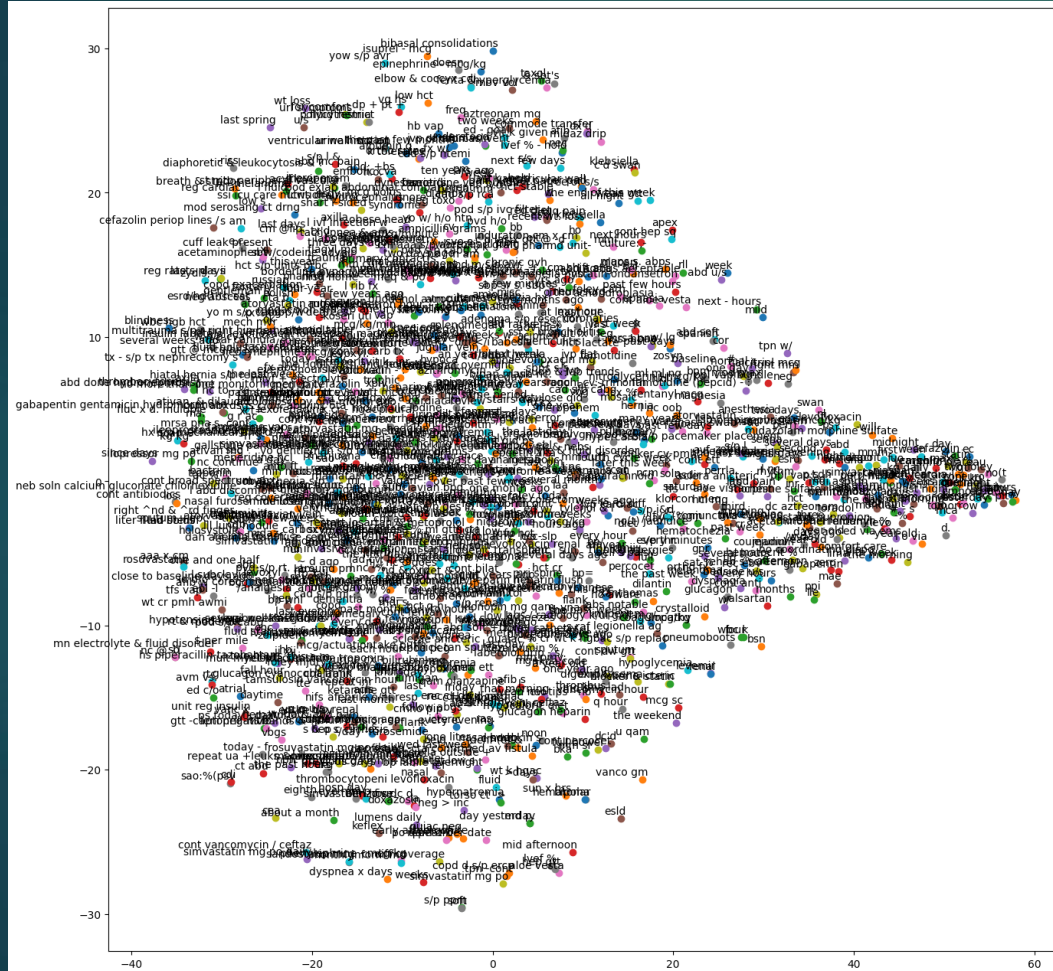
# Word2Vec Assignment

1. Load the CORE untrained model and read the disease\_notes.csv file
2. Build the corpus
3. Create the word2vec embeddings

```
model1.wv['fentanyl']
✓ 0.0s
array([-7.73311080e-03,  1.50146354e-02, -4.68366733e-03,  2.87510501e-03,
        -2.96781072e-03, -3.16113606e-02,  1.30441803e-02,  4.48339768e-02,
        -1.96618997e-02, -1.75622031e-02, -6.35173125e-03, -1.52088255e-02,
        -1.19291870e-02,  1.55969802e-02,  3.40737053e-03, -9.63875372e-03,
         6.23238413e-03, -2.57004127e-02,  2.32704496e-03, -2.75728758e-02,
        -2.24034907e-03,  2.52793636e-03,  1.34125622e-02, -6.96929125e-03,
         1.16599072e-02, -8.63588008e-04, -8.38498678e-03, -1.22807482e-02,
        -2.01310124e-02,  1.38405515e-02,  2.22879481e-02, -4.66979435e-03,
         7.93348532e-03, -2.02539582e-02, -1.04049798e-02,  3.09222117e-02,
        -7.12120067e-03, -6.44101715e-03, -7.68995495e-04, -3.13997380e-02,
        -7.52866827e-03, -2.09959242e-02, -5.51404944e-03,  6.18036417e-03,
         2.29480211e-02, -5.69605688e-03, -5.96011896e-03, -5.96958259e-03,
        -2.97300308e-03,  1.29450690e-02,  1.06825158e-02, -2.08803080e-03,
        -1.27507420e-02, -1.02775702e-02, -2.68260948e-03,  1.29839005e-02,
         1.17887778e-03, -5.85427880e-03, -2.36248542e-02,  1.88551506e-03,
        -6.51479559e-03, -6.04843348e-03,  5.52887749e-03, -1.03047362e-03,
        -1.88045483e-02,  1.73935127e-02,  1.04475860e-02,  1.05643384e-02,
        -1.83757208e-02,  2.49442644e-02, -1.88534555e-04,  9.61792096e-03,
         1.00184474e-02,  7.78188172e-04,  2.67518442e-02,  5.97239472e-03,
        -6.74557668e-05, -9.18865856e-03,  3.99151031e-04,  6.93924446e-03,
        -2.62916391e-03, -3.95182660e-03, -9.78965312e-03,  1.07670231e-02,
         8.99497943e-04, -1.78562221e-03,  1.07684182e-02,  1.00550391e-02,
         3.04169655e-02,  7.53670139e-03,  1.59417484e-02,  5.71035221e-03,
         5.82491932e-03, -6.13961788e-03,  2.71843895e-02,  2.08110046e-02,
         4.93962551e-03, -2.48107668e-02,  1.15699396e-02, -9.98270605e-03],
      dtype=float32)
```

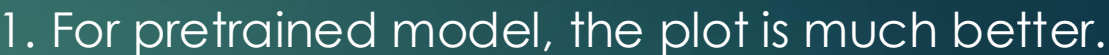


# tSNE Plot for untrained model



1. For untrained model, the tSNE plot is very noisy.





1. For pretrained model, the plot is much better.

# MedSpacy Assignment

1. Install medspacy and load the model.
2. Read disease\_notes.csv after pre-processing
3. Display the pipes in the pipeline

```
import pandas as pd
nlp = spacy.load("en_core_web_sm", disable={"ner"})
nlp = medspacy.load(nlp=nlp)

merged_notes_df = pd.read_csv('disease_notes.csv')
merged_notes_df = preprocessing(merged_notes_df)
notes = merged_notes_df['TEXT']
nlp.pipe_names

✓ 0.3s

['medspacy_pyrush', 'medspacy_target_matcher', 'medspacy_context']
```

4. Add target rules to the medspacy\_target\_matcher pipe
5. Visualize the entities

height: inch cvp: (~)mmhg total in: , ml ml po: tt: , ml ml ivt: ml ml blood products: total out: , ml ml urine: , ml ml ng: stool: drains: balance: - ml ml respiratory support ventilator mode: cmv/assist vt (set): (~) ml vt (spontaneous): (~) ml rr (set): rr (spontaneous): peep: cmho fio: % rsbi deferred: rr > pip: cmho plateau: cmho spo: % abg: /// ve: l/min pao / fio: physical examination peripheral vascular: (right radial pulse: not assessed), (left radial pulse: not assessed), (right dp pulse: not assessed), (left dp pulse: not assessed) skin: not assessed neurologic: responds to: not assessed, movement: not assessed, tone: **not NEGATED\_EXISTENCE** assessed labs / radiology k/ul g/dl mg/dl mg/dl meq/l meq/l mg/dl meq/l meq/l % k/ul : pm : pm : am : am : pm : pm : am : am : pm : am wbc hct plt or tco glucose other labs: pt / ptt / inr://, ck / ckmb / troponin-t://, alt / ast://, alk phos / t bili://, differential-neuts: %, band: %, lymph: %, mono: %, eos: %, d-dimer: ng/ml, fibrinogen: mg/dl, lactic acid: mmol/l, albumin: g/dl, ldh: iu/l, ca++: mg/dl, mg++: mg/dl, po: mg/dl assessment and plan y/o man admitted with respiratory failure to sepsis from mrsa **pna CONDITION** . . # respiratory failure: etiology multifactorial, pt was admitted with influenza c/b secondary mrsa **pna CONDITION** . pt was aggressively volume resuscitated with a depressed ef and continues to moderate bilateral pleural effusions and pulm edema. pt was noted to have inc wob, tachypnea & alkalosis on this am with back up mmv. it is unclear if this is due to agitation vs severely decompensated resp status. rsbi this am, pt **not NEGATED\_EXISTENCE** **likely POSSIBLE\_EXISTENCE** to tolerated extubation, ongoing **family FAMILY** meeting with attg regarding trach placement. - completed course of vancomycin for mrsa **pna CONDITION** , currently day - recheck abg this afternoon, after spending a few hrs on ac - diuresed well overnight | negative, continue with goal of neg cc/day - supportive care, planning for possible trach placement, consulting ethics today . # atrial fibrillation with labile bp: pt with chronic a.fib and labile bps/hr thought due to agitation and extreme sensitivity to medications. pt started on digoxin & metoprolol mg tid, acheived better rate control but still ranging for - - continue metoprolol mg tid - continue digoxin mg iv daily & check am level - continue heparin gtt for anticoag goal ptt - . # **heart failure CONDITION** : pt with h/o diastolic chf, found to have globally depressed ef of -% on admission, possible related to viral myocarditis vs sepsis induced cmp. no rwma,