MIMIC-III NLP

Ankita Savaliya

AI in Healthcare

What Disease Did I Pick?
I selected disease codes related to 4010 – Malignant Essential Hypertension. Malignant essential hypertension is a severe and life-threatening form of high blood pressure that develops rapidly and can cause damage to multiple organs.

What About the Text Data?
The objective of this analysis is to extract medical entities using Named Entity Recognition (NER) with SpaCy, SciSpaCy, Word2Vec, and t-SNE plots.

Additionally, used bc5cdr, BlueBert, MedSpacy to perform a similar analysis.

GitHub and Google Colab Links:

https://colab.research.google.com/github/AnkitaSavaliya/AIH/blob/main/MIMIC_III_NLP.ipynb

https://github.com/AnkitaSavaliya/AIH/blob/main/MIMIC_III_NLP.ipynb

https://github.com/AnkitaSavaliya/AIH/blob/main/MIMIC-III_NLP.pptx

Data Preparation

```python
from google.colab import auth
auth.authenticate_user()
print('Authenticated')

!gcloud projects list

from google.cloud import bigquery

# Construct a BigQuery client object.
client = bigquery.Client(project='clinical-entity-extraction')

"""
ICD codes related to Hypertension:
 4010 - Malignant essential hypertension
 4011 - Benign essential hypertension
 4019 - Unspecified essential hypertension
"""
# Fetch notes only for ICD-9 code 4010(Malignant essential hypertension)
query = """
   SELECT SUBJECT_ID, TEXT, CATEGORY
    FROM `physionet-data.mimiciii_notes.noteevents`
    WHERE SUBJECT_ID IN (
        SELECT d.SUBJECT_ID
        FROM `physionet-data.mimiciii_clinical.diagnoses_icd` d
        WHERE d.ICD9_CODE = '4010' -- Hypertension code
        AND d.SEQ_NUM = 1  -- Assuming 1 indicates primary diagnosis
    )
    AND CATEGORY LIKE 'Discharge summary';
"""

# Run the query
query_job = client.query(query)

# Print the results
noteevents_df = query_job.to_dataframe()

len(noteevents_df)
```

- Fetched rows from noteevents only for ICD-9 CODE **4010** and category '**Discharge Summary**' using the BigQuery client. Here selected records have 4010 (Maligant Hypertension) as primary diagnosis.
- The query returned 162 rows.
- Prepared a DataFrame with the required columns.
- Saved the query result to a CSV/XLSX file to reduce queries to the database.

```python
patients_dict = {"SUBJECT_ID":[],"CATEGORY":[],"TEXT":[]};
for i in range(0, len(noteevents_df)):
   patients_dict["SUBJECT_ID"].append(noteevents_df.loc[i, 'SUBJECT_ID'])
   patients_dict["CATEGORY"].append(noteevents_df.loc[i, 'CATEGORY'])
   patients_dict["TEXT"].append(noteevents_df.loc[i, 'TEXT'])

patients_df = pd.DataFrame(patients_dict)
```

```python
patients_df.shape
```

```
(162, 3)
```

```python
#print first few records
patients_df.head(2)
```

```python
# Download the patients_df dataframe in .csv and excel format
patients_df.to_csv(r'Patient_Summary_4010.csv', index = False)
patients_df.to_excel("Patient_Summary_4010.xlsx")
```

Spacy

Extract and Visualize SpaCy Entities

```
import spacy

# Function to clean and extract tokens
def extract_cleaned_text(text, nlp_model):
    doc = nlp_model(str(text))
    tokens = [token.text for token in doc if not token.is_punct and not token.is_space and not token.is_stop]
    return " ".join(tokens)  # Return cleaned text as a string
```

```
#Load Patient Discharge summary
patients_df_scapy = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/AIH/Patient_Summary_4010.csv")

# Load the spacy model
nlp_spacy = spacy.load('en_core_web_sm')

# Apply token extraction
patients_df_scapy["Processed_Text"] = patients_df_scapy["TEXT"].apply(lambda text: extract_cleaned_text(text, nlp_spacy))
```

```
for i in range(0, 3):
    doc = nlp_spacy( patients_df_scapy['Processed_Text'][i])
    displacy.render(doc, style="ent")
    print("**************************")
```

- Created a common function to clean text for the given model.
- Processed the **Discharge Summary** TEXT column using the **spaCy** model.
- Displayed **spaCy** entities using **displaCy.**

Spacy Entities

Word2Vec and t-SNE Visualization Using SpaCy-Processed Data

```python
def build_corpus(df, model="en_core_web_sm"):
    """
    Extracts named entities from the specified text column in a DataFrame using a spaCy model,
    builds a corpus.

    Parameters:
    - df (pd.DataFrame): DataFrame containing text data.
    - text_column (str): Column name containing processed text.
    - model (str): spaCy model to use (default: "en_core_web_sm").

    Returns:
    - corpus (list of lists): Extracted entities per document.
    """
    nlp = model
    corpus = []

    for _, row in df.iterrows():
        tokens = [ent.text for ent in nlp(row["Processed_Text"]).ents]
        corpus.append(tokens)

    # Calculate word counts
    word_counts = [len(doc) for doc in corpus]

    return corpus
```

```python
#Build corpus for all the notes
corpus_spacy = build_corpus(patients_df_scapy, nlp_spacy)
```

```python
model_word2vec_spacy = Word2Vec(corpus_spacy, min_count=3)
```

```python
model_word2vec_spacy.wv.similar_by_key("BP"), model_word2vec_spacy.wv.similar_by_key("Clonidine")
```

```
([('CT', 0.9998685121536255),
  ('MICU', 0.9998401403427124),
  ('IVC', 0.9998314380645752),
  ('EKG', 0.9998312592506409),
  ('RA', 0.9998170137405396),
  ('IV', 0.9998096227645874),
  ('CXR', 0.9998014569282532),
  ('Family History', 0.9997963309288025),
  ('Known', 0.9997953772544861),
  ('CK', 0.9997934699058533)],
 [('3', 0.9996775388717651),
  ('25', 0.9996731281280518),
  ('100', 0.9996365308761597),
  ('PO', 0.999612033367157),
  ('30', 0.9995716214179993),
  ('50', 0.9995605945587158),
  ('200', 0.9995548725128174),
  ('6', 0.9995482563972473),
  ('7', 0.9995362758636475),
  ('10', 0.9995243549346924)])
```

- Created common function to build corpus using given model SpaCy/SciSpaCy/other

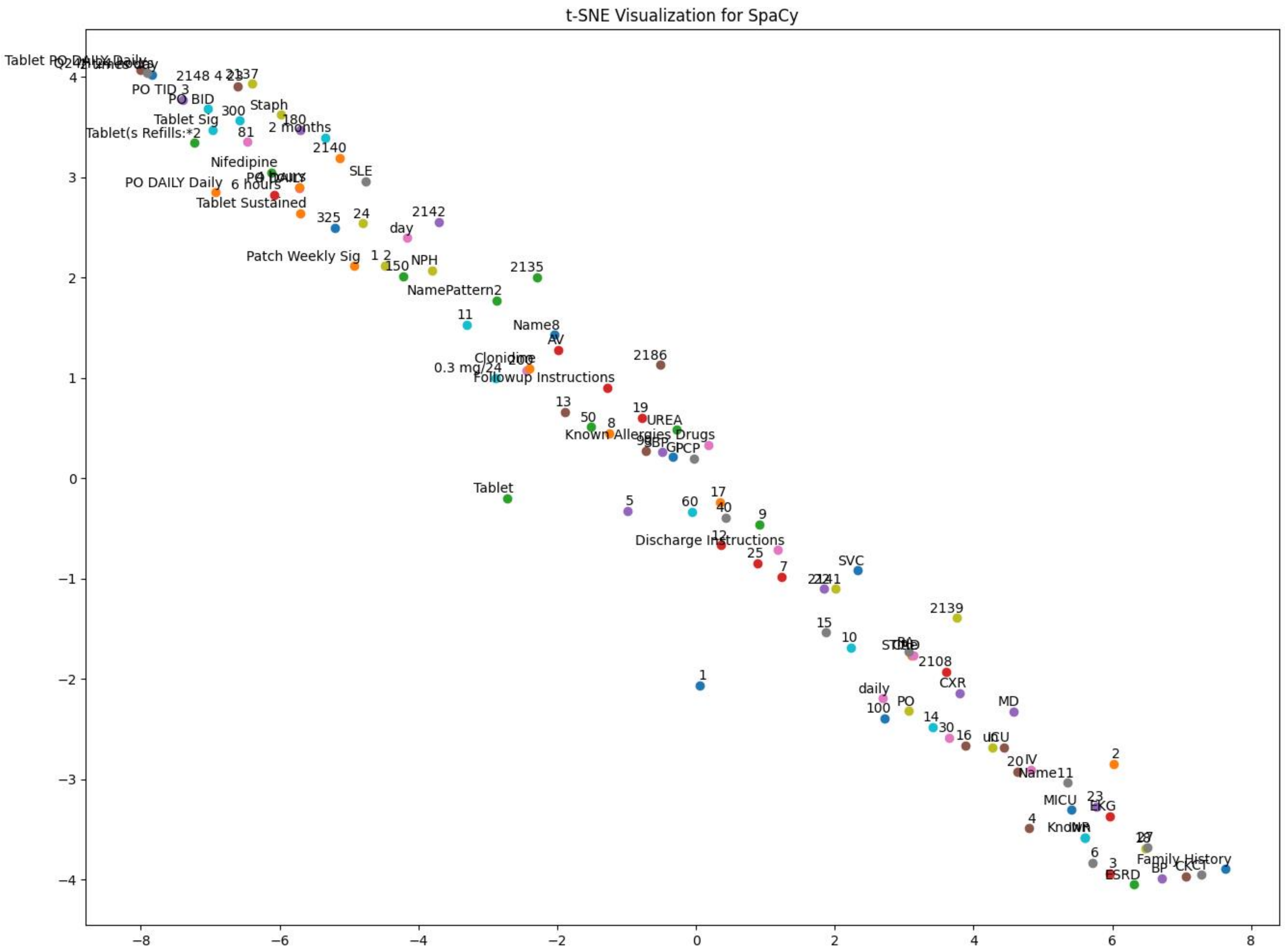
- Defined common function for t-SNE plot.
- Call function using corpus built using Spacy processed text.

```
def tsne_plot(model, words, words_limit = None, model_title="", preTrained=False):
    """
    Creates and displays two t-SNE plots:
    1. Simple scatter plot with labels.
    2. Scatter plot with distance-based coloring.

    Parameters:
    - model: The Word2Vec model or pre-trained model.
    - words: List of words to visualize.
    - words_limit : Limit the number of words to visualize.
    - model_title: Title of the model.
    - preTrained: Boolean flag to choose between Word2Vec or pre-trained model.
    """
    labels = []
    tokens = []

    # Apply t-SNE for dimensionality reduction
    tsne_model = TSNE(perplexity=30, early_exaggeration=12, n_components=2, init='pca', max_iter=1000, random_state=23)

    # Prepare tokens and labels
    for word in words[:words_limit]:
        if preTrained:
            tokens.append(model[word])  # Pre-trained word vectors
        else:
            tokens.append(model.wv[word])  # Word2Vec model vectors
        labels.append(word)

    tokens = np.array(tokens)
    new_values = tsne_model.fit_transform(tokens)
```

```
# First plot: Scatter plot with annotations
plt.figure(figsize=(16,12))
for i in range(len(x)):
    plt.scatter(x[i], y[i])
    plt.annotate(labels[i],
                 xy=(x[i], y[i]),
                 xytext=(5, 2),
                 textcoords='offset points',
                 ha='right',
                 va='bottom')
plt.title(f"t-SNE Visualization for {model_title}")
plt.show()
```

```
tsne_plot(model_word2vec_spacy, np.array(list(model_word2vec_spacy.wv.key_to_index.keys())), 100, 'SpaCy')
```

t-SNE visualization of the top 100 words from Word2Vec (SpaCy), with a limited word count for better label clarity.



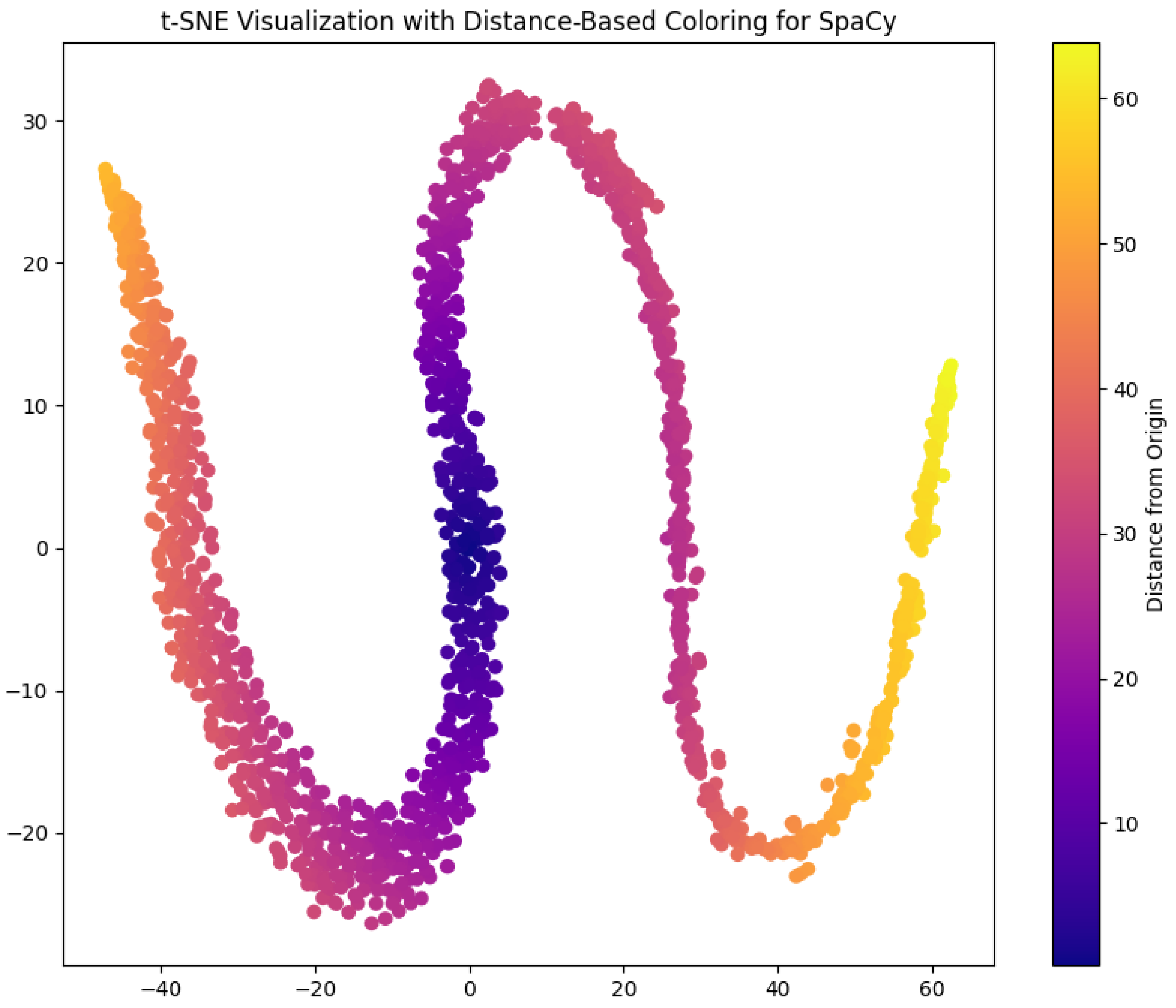
From Word2Vec similarity and above plot we can see that, the entity recognition using SpaCy was limited in extracting hypertension-related terms, likely because it focuses on general English entities rather than clinical ones.

```python
def tsne_plot_no_label(model, words, words_limit=None, model_title="", preTrained=False, reference_word=None):
    """
    Creates and displays a t-SNE plot without labels, using color mapping based on distance.

    Parameters:
    - model: The Word2Vec model or pre-trained model.
    - words: List of words to visualize.
    - words_limit: Maximum number of words to visualize.
    - model_title: Title of the model (used for plot labeling).
    - preTrained: Boolean flag indicating whether to use a pre-trained model.
    - reference_word: (Unused in this function) Placeholder for potential relevance-based coloring.
    """
    tokens = []

    # Apply t-SNE for dimensionality reduction
    tsne_model = TSNE(perplexity=30, early_exaggeration=12, n_components=2, init='pca', max_iter=1000, random_state=23)

    # Extract word vectors
    for word in words[:words_limit]:
        if preTrained:
            tokens.append(model[word])  # Use vectors from a pre-trained model
        else:
            tokens.append(model.wv[word])  # Use vectors from a Word2Vec model

    tokens = np.array(tokens)
    new_values = tsne_model.fit_transform(tokens)

    # Create a scatter plot with color based on distance from the origin
    plt.figure(figsize=(10, 8))
    distances = np.sqrt(new_values[:, 0]**2 + new_values[:, 1]**2)  # Compute Euclidean distance from origin
    plt.scatter(new_values[:, 0], new_values[:, 1], c=distances, cmap='plasma')
    plt.colorbar(label="Distance from Origin")  # Add a color bar for reference
    plt.title(f"t-SNE Visualization with Distance-Based Coloring for {model_title}")
    plt.show()
```

- Defined common function to generates a t-SNE scatter plot of word embeddings, coloring points based on their distance from the origin.
- Call function using corpus built using Spacy processed text.

t-SNE Visualization with Distance-Based Coloring Of All Words from Word2Vec (SpaCy)

```
tsne_plot_no_label(model_word2vec_spacy, np.array(list(model_word2vec_spacy.wv.key_to_index.keys())), None, 'SpaCy')
```



SciSpacy

Extract and Visualize SciSpaCy Entities

```
#Load Patient Discharge summary
patients_df_SciSpaCy = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/AIH/Patient_Summary_4010.csv")

nlp_SciSpaCy = spacy.load('en_core_sci_md')  # Load the specified NLP model
# Apply token extraction
patients_df_SciSpaCy["Processed_Text"] = patients_df_SciSpaCy["TEXT"].apply(lambda text: extract_cleaned_text(text, nlp_SciSpaCy))
```

```
for i in range(0, 3):
    doc = nlp_SciSpaCy( patients_df_SciSpaCy['Processed_Text'][i])
    displacy.render(doc, style="ent", jupyter=True)
    print("****************************************")
```

Admission ENTITY Date 2140 1 19 Discharge Date ENTITY 2140 1 21 Date Birth ENTITY 2117 8 7 Sex F Service ENTITY MEDICINE Allergies Penicillins ENTITY Attending:[**First ENTITY Name3 ENTITY LF 2297 Chief Complaint headache ENTITY Major Surgical Invasive
Procedure Hemodialysis History Present Illness ENTITY Ms. Known lastname ENTITY 22 year old female ENTITY SLE ENTITY lupus nephritis ENTITY ESRD ENTITY HD ENTITY malignant ENTITY HTN ENTITY h/o TTP ENTITY HOCM ENTITY presents HA
hypertensive ENTITY urgency ENTITY Awoke ENTITY a.m. 8/10 left sided frontal HA sure d/t ENTITY flare uveitis ENTITY started Monday ENTITY d/t HTN ENTITY Decided skip ENTITY HD ENTITY come ED ENTITY evaluation ENTITY vision changes ENTITY
numbness weakness ENTITY change gait chest ENTITY pain SOB ENTITY + Diarrhea ENTITY x 1 day ENTITY ED ENTITY patient ENTITY 217/140 elevated ENTITY 254/152 > received labetolol ENTITY IV 30 mg x 1 MSO4 ENTITY 4 mg pressures ENTITY
dropped SBPs ENTITY 208 HA ENTITY improved Repeat labetolol ENTITY 50 mg x 1 repeated dose ENTITY morphine ENTITY dropped pressures 193/134 > labetolol ENTITY gtt ENTITY started asa given HA ENTITY resolved Head CT ENTITY negative ENTITY
intracranial bleed ENTITY CXR ENTITY unremarkable ROS cold ENTITY past week fevers chills ENTITY CP ENTITY SOB ENTITY N/V ENTITY + diarrhea ENTITY arrival ENTITY floor patient BP ENTITY 191/126 labetolol ENTITY gtt ENTITY started sxs HA
states compliant meds ENTITY mother ENTITY cooks salt ENTITY adherent ENTITY diet ENTITY Past Medical History 1 Lupus 2134 ENTITY Diagnosed ENTITY began swolen fingers rash painful joints 2 ENTITY ESRD ENTITY secodary ENTITY SLE ENTITY 2135
initially cytoxan 1 dose ENTITY 3 months ENTITY 2 years began dialysis ENTITY 3 times week ENTITY 2137 T Th Sat Awaiting ENTITY living donor transplant ENTITY mother 3 HTN ENTITY 2137 Normal BPs ENTITY run 180's/120 1 hypertensive crisis ENTITY
precipitated ENTITY seizures ENTITY past 4 Uveitis ENTITY secondary ENTITY SLE ENTITY 4 15 5 HOCM ENTITY Echo ENTITY 2137 6 Vaginal bleeding ENTITY 2139 9 20 7 Mulitple episodes dialysis reactions ENTITY 8 Anemia ENTITY 9 Coag neg ENTITY
Staph bacteremia ENTITY HD line infection ENTITY 6 15 10 H/O UE ENTITY clot ENTITY coumadin ENTITY longer Social History ENTITY Lives Location ENTITY 669 mother ENTITY 16 year ENTITY old brother ENTITY Graduated Name2 NI School ENTITY got
sick ENTITY currently working ENTITY attending school Denies ENTITY T/E/D. Family History ENTITY -No history ENTITY SLE ENTITY -Grandfather HTN ENTITY -Distant history DM ENTITY -No history clotting disorders ENTITY -No history ENTITY autoimmune
diseases ENTITY Physical Exam Vitals ENTITY 98.0 173/51 86 15 100 RA ENTITY HEENT ENTITY L eye injected w/periorbital edema R eye reactive w/ EOMI anicteric ENTITY sclera ENTITY MMM ENTITY OP ENTITY clear Neck supple ENTITY LAD thyromegaly Cardiac
RRR ENTITY NL S1 S2 ENTITY + S4 III/VI ENTITY systolic ejection murmur LUSB ENTITY radiating apex axilla ENTITY intensifies w/ Valsalva rub ENTITY Lungs CTAB ENTITY wheezes ENTITY rhonchi crackles ENTITY Abd soft NTND NABS ENTITY HSM ENTITY
rebound guarding ENTITY GU ENTITY CVAT ENTITY Ext warm 2 ENTITY + DP ENTITY pulses ENTITY C/C/E ENTITY L femoral dialysis catheter ENTITY Neuro AOx3 CN II-XII ENTITY intact strength/sensation ENTITY grossly intact Pertinent Results UA ENTITY
mod bld 100 protein ENTITY present prior UAs ENTITY Radiology ENTITY CXR ENTITY acute CP ENTITY abnormality ENTITY EKG ENTITY NSR ENTITY nml ENTITY axis nml intervals borderline LAE LVH J point elevation V2,V3 TWI ENTITY aVL V5 V6 change
compared prior 2139 11 26 CT ENTITY HEAD intracranial hemorrhage ENTITY Brief Hospital ENTITY Course A/P ENTITY Patient ENTITY 22 year old female ENTITY SLE ENTITY lupus nephritis ENTITY ESRD ENTITY HD ENTITY presents hypertensive ENTITY
urgency Hypertensive ENTITY urgency Unclear precipitant ENTITY Possibly secondary pain ENTITY worsening ENTITY uveitis ENTITY Compliant ENTITY meds Denies ENTITY illicits ENTITY tox screen ENTITY negative ENTITY Patient ENTITY started labetolol
ENTITY drip ED ENTITY good BP response ENTITY subsequently transitioned PO ENTITY anti-hypertensives ICU ENTITY maintenance ENTITY stable SBPs ENTITY 150s-170s baseline ENTITY 170s-190s nephrologist ENTITY recommendations home lisinopril ENTITY
increased ENTITY 40 mg po bid ENTITY 40 mg po qd better baseline ENTITY BP ENTITY control clinical evidence ENTITY end organ damage UA ENTITY difficult ro interpret setting CRF ENTITY CE ENTITY x 1 negative ENTITY Headache ENTITY evidence ENTITY
CT ENTITY intracranial bleed ENTITY Headaches controlled ENTITY morphine sulfate ENTITY resolved time discharge Uveitis ENTITY Followed outpatient ENTITY optho specialist ENTITY Optho consulted ENTITY patient ENTITY request ESRD ENTITY Secondary

Word2Vec and t-SNE Visualization Using SciSpaCy-Processed Data

```
corpus_scispacy = build_corpus(patients_df_SciSpaCy, nlp_SciSpaCy)
```

```
model_word2vec_scispacy = Word2Vec(corpus_scispacy, min_count=3)
```

```
model_word2vec_scispacy.wv.similar_by_key("BP"), model_word2vec_scispacy.wv.similar_by_key("Clonidine")
```

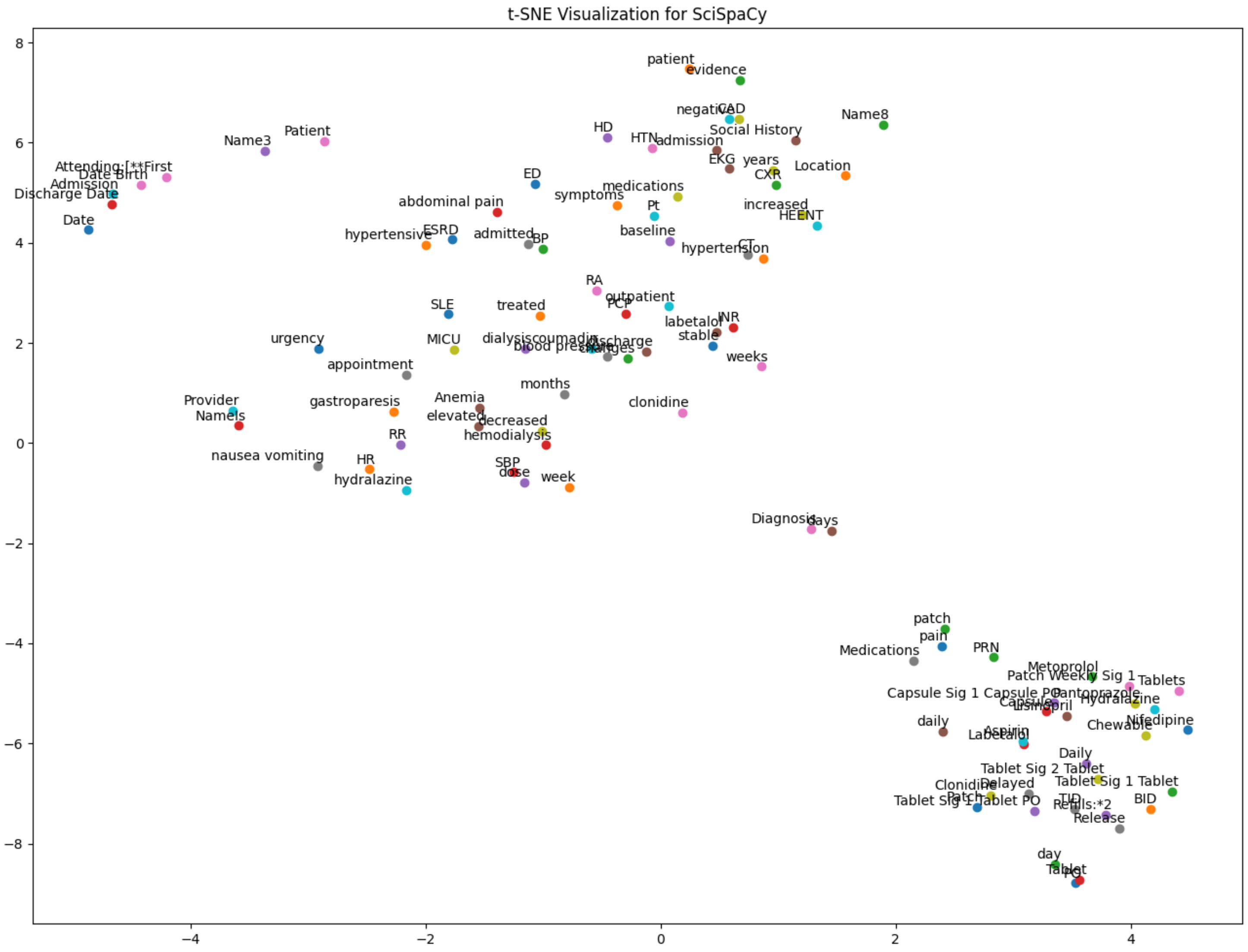
```
([('HR', 0.9993711709976196),
  ('admitted', 0.9993606805801392),
  ('MICU', 0.9992586374282837),
  ('RR', 0.9992402791976929),
  ('gastroparesis', 0.9992167353630066),
  ('patient', 0.9991229176521301),
  ('ED', 0.9990416169166565),
  ('symptoms', 0.9990056753158569),
  ('HTN', 0.998991072177887),
  ('HD', 0.9989602565765381)],
 [('Patch', 0.9993199706077576),
  ('Lisinopril', 0.998427152633667),
  ('Patch Weekly Sig 1', 0.9981398582458496),
  ('Prednisone', 0.997868537902832),
  ('Transdermal QWED', 0.9976481795310974),
  ('Labetalol', 0.9975727200508118),
  ('Aspirin', 0.9974137544631958),
  ('Amlodipine', 0.9971833229064941),
  ('Metoclopramide', 0.9970840811729431),
  ('Refills:*0', 0.99689781665802)])
```

```
len(model_word2vec_scispacy.wv.key_to_index.keys())
```

```
3276
```

```
tsne_plot(model_word2vec_scispacy, np.array(list(model_word2vec_scispacy.wv.key_to_index.keys())), 100, 'SciSpaCy')
```

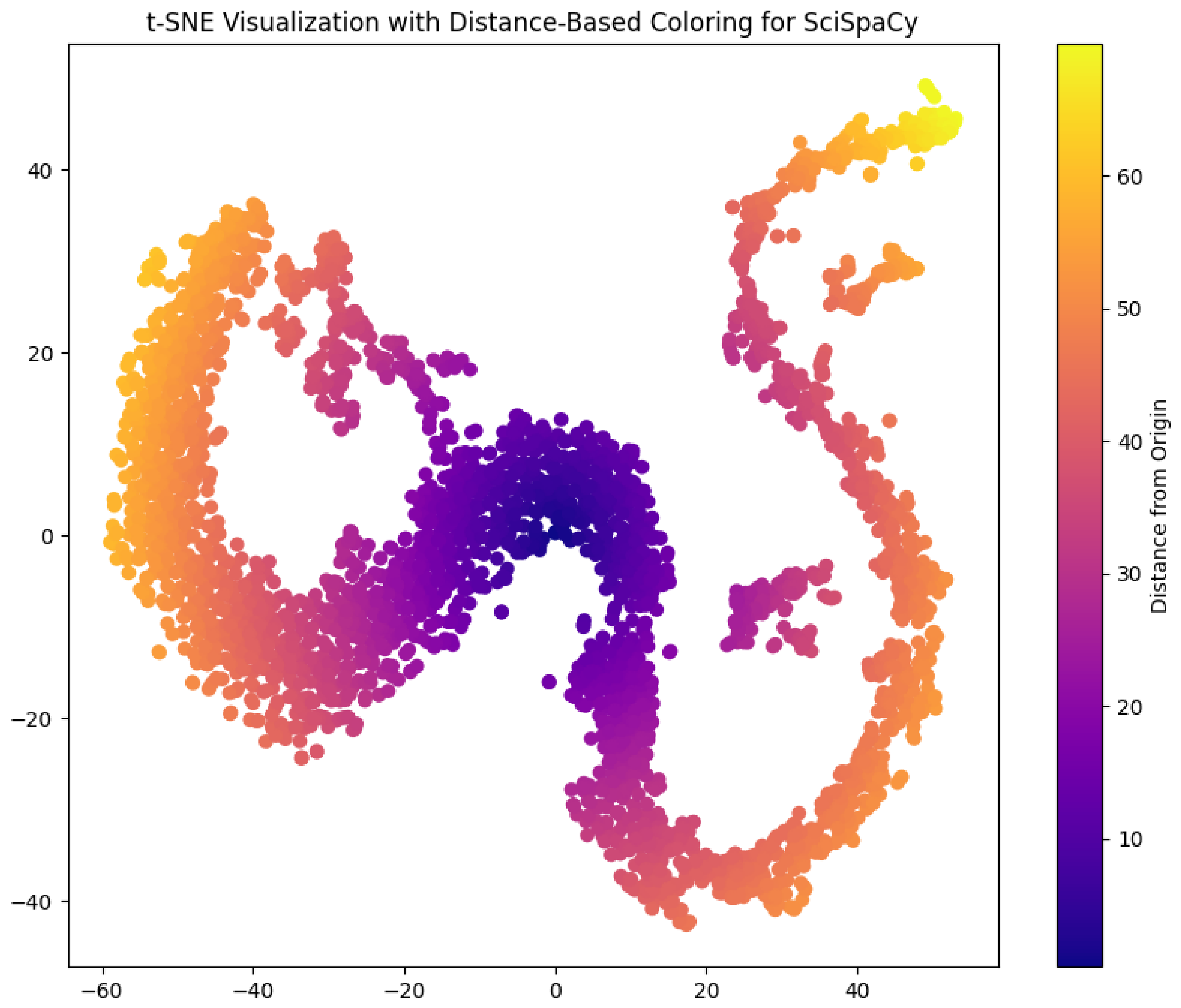
t-SNE visualization of the top 100 words from Word2Vec (SciSpaCy), with a limited word count for better label clarity.



From Word2Vec similarity and above plot , SciSpaCy primarily recognized medication names and formulations, such as Clonidine and Labetalol, but it did not specifically highlight key hypertension-related entities beyond drug mentions.

t-SNE Visualization with Distance-Based Coloring Of All Words from Word2Vec (SciSpaCy)

```
tsne_plot_no_label(model_word2vec_scispacy, np.array(list(model_word2vec_scispacy.wv.key_to_index.keys())), None, 'SciSpaCy')
```



BC5CDR (BioCreative V Chemical-Disease Relation)

BC5CDR Entity Visualization Using SciSpaCy-Processed Data

```
nlp__bc5cdr = en_ner_bc5cdr_md.load()

# Note: Displaying all data points is causing the notebook to become oversized, unable to check-in
# Visualize named entities using displacy for first few notes
for i in range(0, 3):
    doc = nlp__bc5cdr( patients_df_SciSpaCy['Processed_Text'][i])
    displacy.render(doc, style="ent", jupyter=True)
    print("*****************************************")
```

en_ner_bc5cdr_md is a Named Entity Recognition (NER) model from SciSpaCy that specializes in identifying **diseases** and **chemicals** in text

Admission Date 2140 1 19 Discharge Date 2140 1 21 Date Birth 2117 8 7 Sex F Service MEDICINE Allergies Penicillins CHEMICAL Attending:[**First Name3 LF 2297 Chief Complaint headache DISEASE Major Surgical Invasive Procedure Hemodialysis History Present Illness Ms. Known lastname 22 year old female SLE lupus nephritis ESRD HD malignant HTN DISEASE h/o TTP HOCM DISEASE presents HA hypertensive DISEASE urgency Awoke a.m. 8/10 left sided frontal HA sure d/t flare uveitis DISEASE started Monday d/t HTN DISEASE Decided skip HD come ED evaluation vision changes numbness weakness DISEASE change gait chest pain DISEASE SOB + Diarrhea DISEASE x 1 day ED patient 217/140 elevated 254/152 > received labetolol CHEMICAL IV 30 mg x 1 MSO4 4 mg pressures dropped SBPs 208 HA improved Repeat labetolol CHEMICAL 50 mg x 1 repeated dose morphine CHEMICAL dropped pressures 193/134 > labetolol CHEMICAL gtt started asa given HA resolved Head CT negative intracranial bleed DISEASE CXR unremarkable ROS cold past week fevers chills CP SOB N/V + DISEASE diarrhea DISEASE arrival floor patient BP 191/126 labetolol CHEMICAL gtt started sxs HA states compliant meds mother cooks salt adherent diet Past Medical History 1 Lupus 2134 Diagnosed began swolen fingers rash DISEASE painful joints 2 ESRD DISEASE secodary SLE 2135 initially cytoxan CHEMICAL 1 dose 3 months 2 years began dialysis 3 times week 2137 T Th Sat Awaiting living donor transplant mother 3 HTN DISEASE 2137 Normal BPs run 180's/120 1 hypertensive DISEASE crisis precipitated seizures DISEASE past 4 Uveitis DISEASE secondary SLE 4 15 5 HOCM DISEASE Echo 2137 6 Vaginal bleeding DISEASE 2139 9 20 7 Mulitple episodes dialysis reactions 8 Anemia DISEASE 9 Coag neg Staph bacteremia DISEASE HD line infection DISEASE 6 15 10 H/O UE clot coumadin CHEMICAL longer Social History Lives Location 669 mother 16 year old brother Graduated Name2 NI School got sick currently working attending school Denies T/E/D. Family History -No history SLE -Grandfather HTN DISEASE -Distant history DM CHEMICAL -No history clotting disorders -No history DISEASE autoimmune diseases DISEASE Physical Exam Vitals 98.0 173/51 86 15 100 RA HEENT L eye injected w/periorbital edema R eye reactive w/ DISEASE EOMI anicteric sclera MMM OP clear Neck supple LAD thyromegaly DISEASE Cardiac RRR NL S1 S2 + S4 III/VI systolic ejection murmur LUSB radiating apex axilla intensifies w/ Valsalva rub Lungs CTAB wheezes rhonchi CHEMICAL crackles Abd soft NTND NABS DISEASE HSM rebound guarding GU CVAT Ext warm 2 + DP pulses C/C/E L femoral dialysis catheter Neuro AOx3 CN II-XII intact strength/sensation grossly intact Pertinent Results UA CHEMICAL mod bld 100 protein present prior UAs CHEMICAL Radiology CXR acute CP abnormality CHEMICAL EKG NSR DISEASE nml axis nml intervals borderline LAE LVH J DISEASE point elevation V2,V3 TWI aVL V5 V6 change compared prior 2139 11 26 CT HEAD intracranial hemorrhage DISEASE Brief Hospital Course A/P Patient 22 year old female SLE lupus nephritis ESRD DISEASE HD presents hypertensive DISEASE urgency Hypertensive DISEASE urgency Unclear precipitant Possibly secondary pain DISEASE worsening uveitis DISEASE Compliant meds Denies illicits tox screen negative Patient started labetolol CHEMICAL drip ED good BP response subsequently transitioned PO anti-hypertensives ICU maintenance stable SBPs 150s-170s baseline 170s-190s nephrologist recommendations home lisinopril CHEMICAL increased 40 mg po bid 40 mg po qd better baseline BP control clinical evidence end organ damage UA CHEMICAL difficult ro interpret setting CRF DISEASE CE x 1 negative Headache DISEASE evidence CT intracranial bleed DISEASE Headaches DISEASE controlled morphine CHEMICAL sulfate resolved time discharge Uveitis DISEASE Followed outpatient optho specialist Optho consulted patient request ESRD DISEASE Secondary lupus nephritis DISEASE transplant list Patient received hemodialysis house 500 ml ultrafiltrate complications dry weight 45 kg patient Began Sevalamer 800 TID meals Given difficulty interpreting renin aldosterone CHEMICAL levels acutely ill DISEASE patients drawn need drawn outpatient follow Medications Admission Lisinopril CHEMICAL 40 mg PO QD Labetalol CHEMICAL 600 PO TID Valsartan CHEMICAL 320 mg PO QD Clonidine CHEMICAL 0.3 mg transdermal QW Prednisone CHEMICAL 40 mg PO QD Atropine CHEMICAL 1 Hospital1 Prednisolone Acetate CHEMICAL 1 Q1H Moxifloxacin CHEMICAL eye drops qid Lorazepam CHEMICAL 1 mg PO Q4 6H PRN Discharge Medications 1

Word2Vec and t-SNE Visualization For BC5CDR

```
corpus_bc5cdr = build_corpus(patients_df_SciSpaCy, nlp__bc5cdr)
```

```
model_word2vec_bc5cdr = Word2Vec(corpus_bc5cdr, min_count=3, window=2, vector_size=100)
```

```
model_word2vec_bc5cdr.wv.similar_by_word("BP"), model_word2vec_bc5cdr.wv.similar_by_word("Clonidine")
```

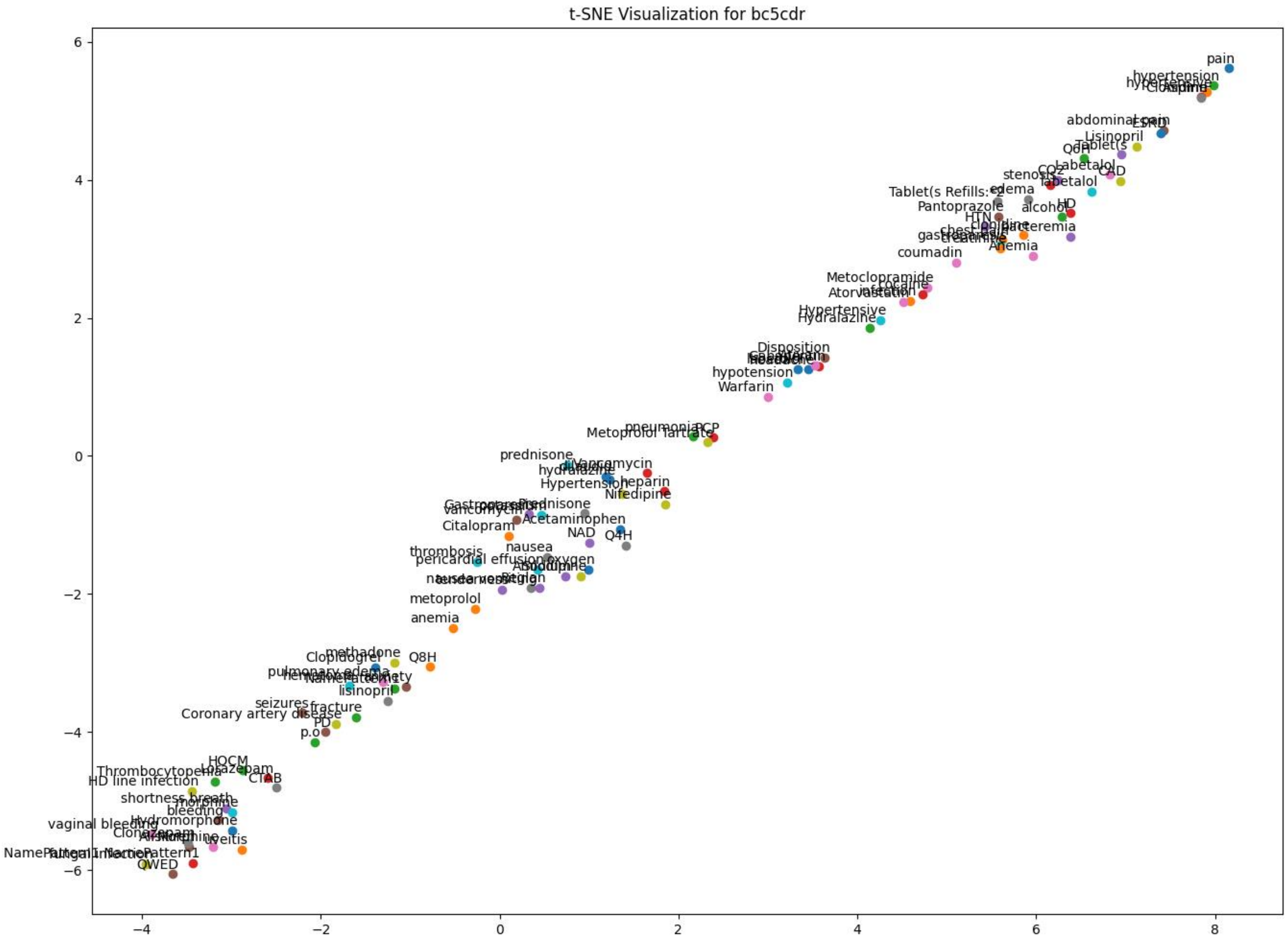
```
([('qid', 0.7542693614959717),
  ('Lactate', 0.7493560910224915),
  ('CP', 0.7465789914131165),
  ('papilledema', 0.745745837688446),
  ('atrial fibrillation', 0.7431225180625916),
  ('reglan', 0.7429000735282898),
  ('pleural effusions', 0.7415630221366882),
  ('Carvedilol', 0.740106999874115),
  ('fentanyl', 0.7399438619613647),
  ('coronary artery disease', 0.7399159073829651)],
 [('hypertension', 0.9987433552742004),
  ('pain', 0.9987409114837646),
  ('Aspirin', 0.998708963394165),
  ('Labetalol', 0.9986240267753601),
  ('hypertensive', 0.9985849857330322),
  ('abdominal pain', 0.9985615611076355),
  ('Lisinopril', 0.9985013604164124),
  ('chest pain', 0.9984785914421082),
  ('HTN', 0.9983934760093689),
  ('ESRD', 0.9983604550361633)])
```

```
len(model_word2vec_bc5cdr.wv.key_to_index.keys())
```

```
885
```

```
tsne_plot(model_word2vec_bc5cdr,np.array(list(model_word2vec_bc5cdr.wv.key_to_index.keys())), 100, 'bc5cdr')
```

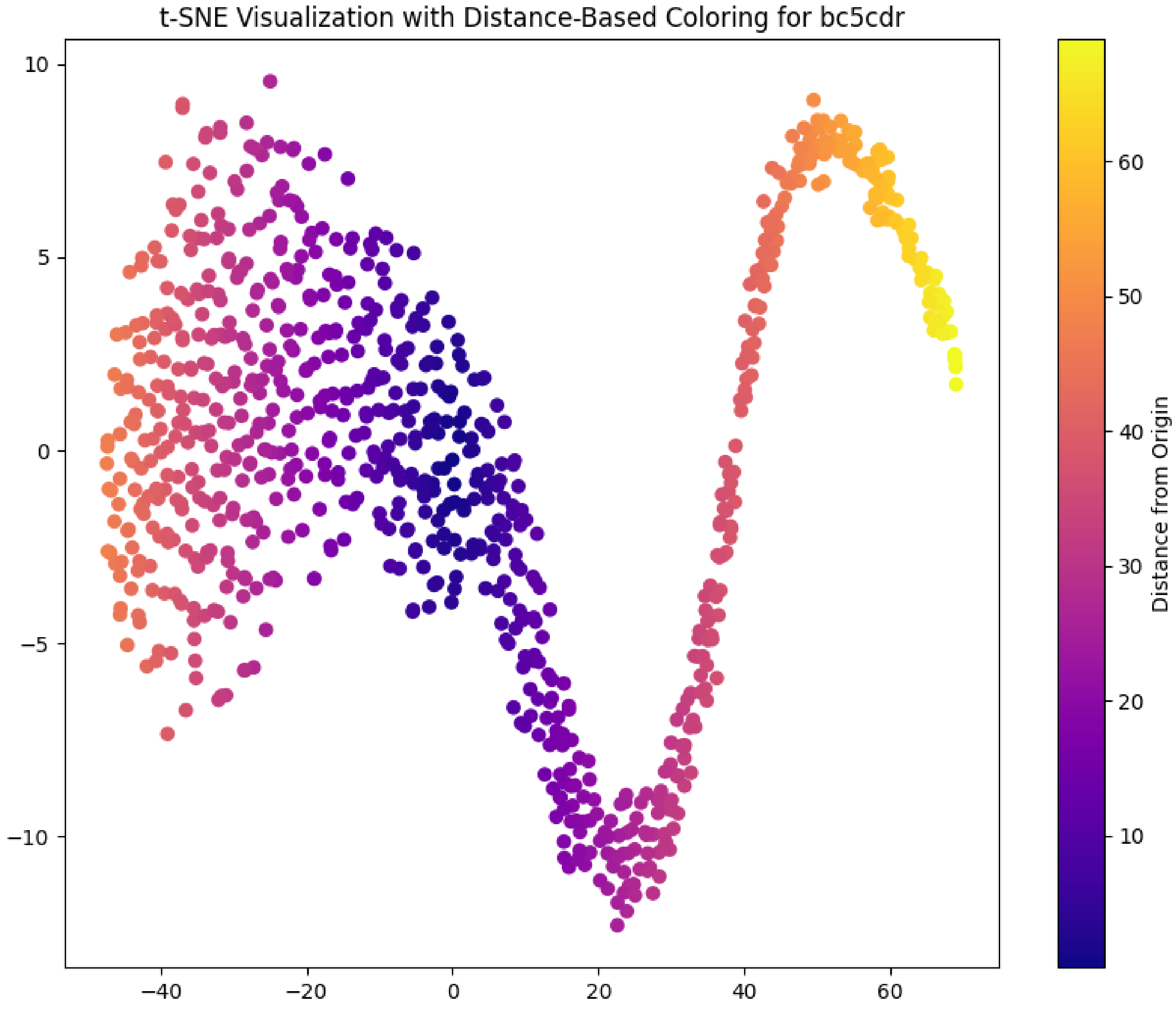
t-SNE visualization of the top 100 words from Word2Vec (bc5cdr), with a limited word count for better label clarity.



Based on Word2Vec similarity and the above plot, BC5CDR appears to capture disease and medication entities well, with a good emphasis on hypertension-related terms (e.g., "hypertension," "hypertensive," "lisinopril" , "metoprolol").

t-SNE Visualization with Distance-Based Coloring Of All Words from Word2Vec (bc5cdr)

```
tsne_plot_no_label(model_word2vec_bc5cdr,np.array(list(model_word2vec_bc5cdr.wv.key_to_index.keys())), None, 'bc5cdr')
```



MedSpacy

Custom Rule-Based Entity Extraction with MedspaCy NLP Pipeline

```python
# Load MedspaCy NLP pipeline
nlp_medspacy = medspacy.load()

# Add rules for target concept extraction
target_matcher = nlp_medspacy.get_pipe("medspacy_target_matcher")
# Define custom rules for better entity detection
target_rules = [
    TargetRule("hyperlipidemia", "DISEASE"),
    TargetRule("O2", "SUBSTANCE"),
    TargetRule("FiO2", "SUBSTANCE"),
    TargetRule("hypertension", "DISEASE"),
    TargetRule("obesity", "CONDITION"),
    TargetRule("cardiac", "CONDITION"),
    TargetRule("SLE", "DISEASE"),  # Systemic Lupus Erythematosus
    TargetRule("lupus nephritis", "DISEASE"),
    TargetRule("ESRD", "DISEASE"),  # End-Stage Renal Disease
    TargetRule("dialysis", "TREATMENT"),  # Hemodialysis is also treatment
    TargetRule("hemodialysis", "TREATMENT"),
    TargetRule("SBP", "MEASUREMENT"),  # Systolic Blood Pressure
    TargetRule("HR", "MEASUREMENT"),  # Heart Rate
    TargetRule("TPN", "TREATMENT"),
    TargetRule("Prednisone", "MEDICATION"),
    TargetRule("Lisinopril", "MEDICATION"),
    TargetRule("Labetalol", "MEDICATION"),
    TargetRule("Clonidine", "MEDICATION"),
    TargetRule("Valsartan", "MEDICATION"),
    TargetRule("Sevelamer", "MEDICATION"),
    TargetRule("Atropine", "MEDICATION"),
    TargetRule("Morphine sulfate", "MEDICATION"),
    TargetRule("Diarrhea", "SYMPTOM"),
    TargetRule("Headache", "SYMPTOM"),
    TargetRule("nausea", "SYMPTOM"),
    TargetRule("vomiting", "SYMPTOM"),
    TargetRule("shortness of breath", "SYMPTOM"),
    TargetRule("fever", "SYMPTOM"),
    TargetRule("chills", "SYMPTOM")
]

target_matcher.add(target_rules)
```

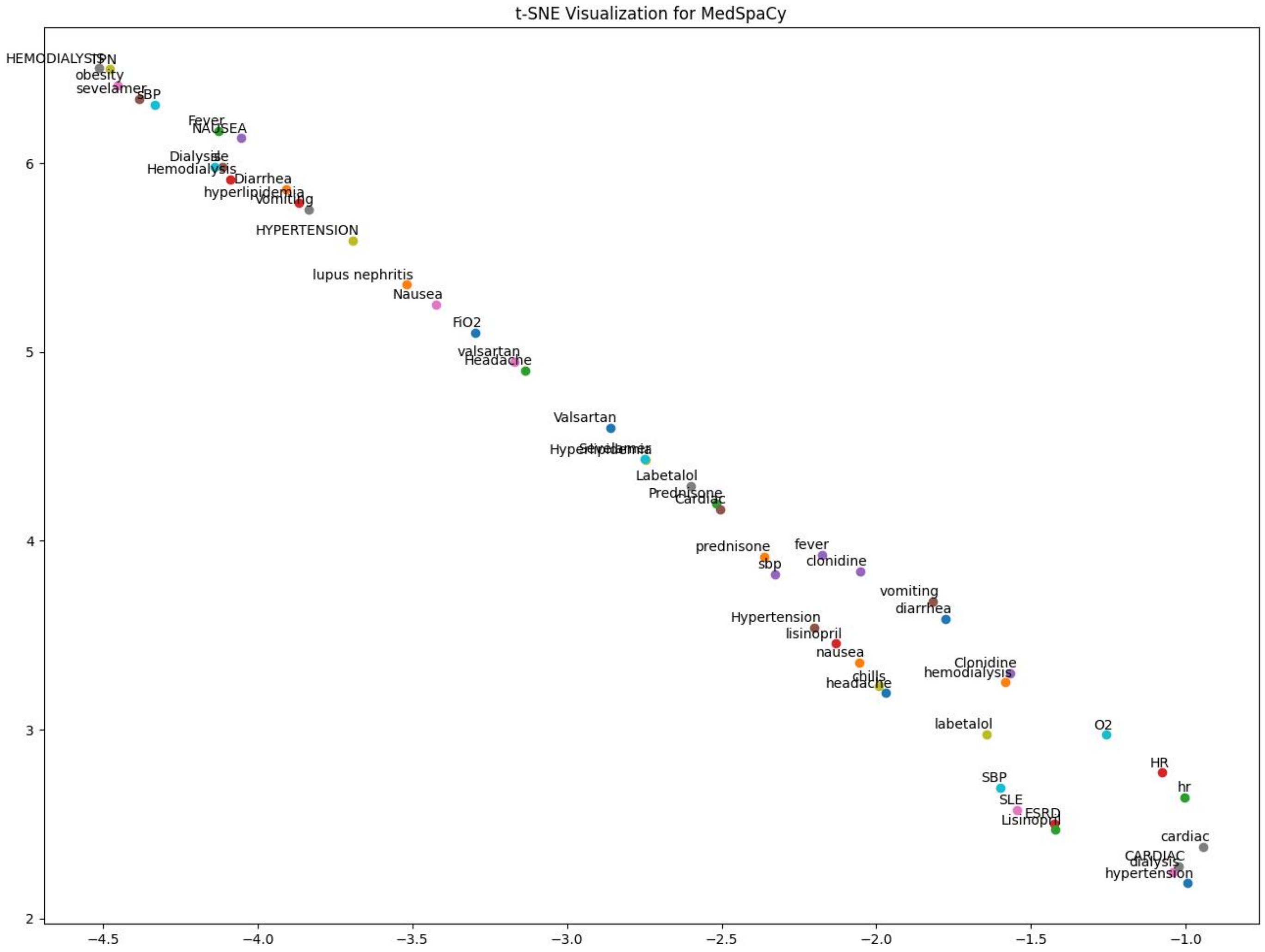
- MedSpaCy is a library designed for processing clinical and biomedical text.
- In this code, MedSpaCy is being enhanced by adding custom target rules to better detect specific medical entities such as diseases, treatments, symptoms, and medications in clinical notes.
- Loaded the MedspaCy NLP pipeline. Used the medspacy_target_matcher to add custom rules for extracting medical concepts.
- Defined specific target rules to identify entities like diseases (e.g., hypertension), treatments (e.g., hemodialysis), medications (e.g., Lisinopril), symptoms (e.g., headache), and measurements (e.g., SBP).
- Applied these rules to clinical text for improved entity detection.

MedSpacy Visualization Using SciSpaCy-Processed Data

```
# Visualize named entities using displacy for first few notes
for i in range(0, 3):
  # Process the shift note
  doc = nlp_medspacy(patients_df_SciSpaCy['Processed_Text'][i])
  # visulize
  visualize_ent(doc)
  print("*****************************")
```

Admission Date 2140 1 19 Discharge Date 2140 1 21 Date Birth 2117 8 7 Sex F Service MEDICINE Allergies Penicillins Attending:[**First Name3 LF 2297 Chief Complaint headache SYMPTOM Major Surgical Invasive Procedure Hemodialysis TREATMENT History Present Illness Ms. Known lastname 22 year old female SLE DISEASE lupus nephritis DISEASE ESRD DISEASE HD malignant HTN h/o TTP HOCM presents HA hypertensive urgency Awoke a.m. 8/10 left sided frontal HA sure d/t flare uveitis started Monday d/t HTN Decided skip HD come ED evaluation vision changes numbness weakness change gait chest pain SOB + Diarrhea SYMPTOM x 1 day ED patient 217/140 elevated 254/152 > received labetolol IV 30 mg x 1 MSO4 4 mg pressures dropped SBPs 208 HA improved Repeat labetolol 50 mg x 1 repeated dose morphine dropped pressures 193/134 > labetolol gtt started asa given HA resolved NEGATED_EXISTENCE Head CT negative intracranial bleed CXR unremarkable ROS cold past week fevers chills SYMPTOM CP SOB N/V + diarrhea SYMPTOM arrival floor patient BP 191/126 labetolol gtt started sxs HA states compliant meds mother FAMILY cooks salt adherent diet Past Medical History HISTORICAL 1 Lupus 2134 Diagnosed began swolen fingers rash painful joints 2 ESRD DISEASE secodary SLE DISEASE 2135 initially cytoxan 1 dose 3 months 2 years began dialysis TREATMENT 3 times week 2137 T Th Sat Awaiting living donor transplant mother 3 HTN 2137 Normal BPs run 180's/120 1 hypertensive crisis precipitated seizures past 4 Uveitis secondary SLE DISEASE 4 15 5 HOCM Echo 2137 6 Vaginal bleeding 2139 9 20 7 Mulitple episodes dialysis TREATMENT reactions 8 Anemia 9 Coag neg Staph bacteremia HD line infection 6 15 10 H/O UE clot coumadin longer Social History Lives Location 669 mother 16 year old brother Graduated Name2 NI School got sick currently working attending school Denies NEGATED_EXISTENCE T/E/D. Family FAMILY History -No history HISTORICAL SLE DISEASE - Grandfather HTN -Distant history DM -No history clotting disorders -No history HISTORICAL autoimmune diseases Physical Exam Vitals 98.0 173/51 86 15 100 RA HEENT L eye injected w/periorbital edema R eye reactive w/ EOMI anicteric sclera MMM OP clear Neck supple LAD thyromegaly Cardiac CONDITION RRR NL S1 S2 + S4 III/VI systolic ejection murmur LUSB radiating apex axilla intensifies w/ Valsalva rub Lungs CTAB wheezes rhonchi crackles Abd soft NTND NABS HSM rebound guarding GU CVAT Ext warm 2 + DP pulses C/C/E L femoral dialysis TREATMENT catheter Neuro AOx3 CN II-XII intact strength/sensation grossly intact Pertinent Results UA mod bld 100 protein present prior UAs Radiology CXR acute CP abnormality EKG NSR nml axis nml intervals borderline LAE LVH J point elevation V2,V3 TWI aVL V5 V6 change compared prior 2139 11 26 CT HEAD intracranial hemorrhage Brief Hospital Course A/P Patient 22 year old female SLE DISEASE lupus nephritis DISEASE ESRD DISEASE HD presents hypertensive urgency Hypertensive urgency Unclear precipitant Possibly secondary pain worsening uveitis Compliant meds Denies NEGATED_EXISTENCE illicits tox screen negative Patient started labetolol drip ED good BP response subsequently transitioned PO anti-hypertensives ICU maintenance stable SBPs 150s-170s baseline 170s-190s nephrologist recommendations home lisinopril MEDICATION increased 40 mg po bid 40 mg po qd better baseline BP control clinical evidence end organ damage UA difficult ro POSSIBLE_EXISTENCE interpret setting CRF CE x 1 negative Headache SYMPTOM evidence CT intracranial bleed Headaches controlled morphine sulfate MEDICATION resolved NEGATED_EXISTENCE time discharge Uveitis Followed outpatient optho specialist Optho consulted patient request ESRD DISEASE Secondary lupus nephritis DISEASE transplant list Patient received hemodialysis TREATMENT house 500 ml ultrafiltrate complications dry weight 45 kg patient Began Sevalamer 800 TID meals Given difficulty interpreting renin aldosterone levels acutely ill patients drawn need drawn outpatient follow Medications Admission Lisinopril MEDICATION 40 mg PO QD Labetalol MEDICATION 600 PO TID Valsartan MEDICATION 320 mg PO QD Clonidine MEDICATION 0.3 mg transdermal QW Prednisone MEDICATION 40 mg PO QD Atropine MEDICATION 1 Hospital1 Prednisolone Acetate 1 Q1H Moxifloxacin eye drops qid Lorazepam 1 mg PO Q4 6H PRN Discharge Medications 1 Labetalol MEDICATION 200 mg Tablet Sig 3 Tablet PO TID 3 times day Tablet(s 2 Clonidine MEDICATION 0.3 mg/24 hr MEASUREMENT Patch Weekly Sig 1 Patch Weekly Transdermal QTHUR Thursday 3 Atropine MEDICATION 1 Drops Sig 1 Drop Ophthalmic Hospital1 2 times day 4 Lorazepam 1 mg Tablet Sig 1 Tablet PO Q4 6H 4 6 hours needed 5 Valsartan MEDICATION 160 mg Tablet Sig 2 Tablet PO DAILY Daily 6 Prednisolone Acetate 1 Drops Suspension Sig 1 Drop Ophthalmic Q1H hour 7 Lisinopril MEDICATION 40 mg Tablet Sig 1 Tablet PO twice day Disp:*60 Tablet(s Refills:*2 8 Sevelamer MEDICATION 800 mg Tablet Sig 1 Tablet PO TID 3 times day Disp:*90 Tablet(s Refills:*2 9 Prednisone MEDICATION 20 mg Tablet Sig 2 Tablet PO day 10 Blood Pressure Kit Kit Sig 1 Kit Miscellaneous day Disp:*1 Kit Refills:*0 Discharge Disposition Home Discharge Diagnosis Hypertensive urgency Discharge Condition Good Discharge Instructions blood pressure medications prescribed adhere low-salt diet increased levels sodium drive blood pressure discharged prescription home blood pressure monitor use daily measurements primary care physician Initial PRE systolic blood pressures greater 180 experience headaches nausea SYMPTOM vomiting SYMPTOM chest pain shortness breath concerning symptoms Followup Instructions resume hemodialysis TREATMENT according regular schedule scheduled Dr. Name8 NamePattern2 NamePattern1 4883 Division Nephrology Wednesday 2 3 9:30 Telephone/Fax 1 435 need reschedule scheduled follow-up primary care physician NamePattern4 Name4 NamePattern1 NamePattern1 2423 Tuesday 1 26 3:30 PM Telephone/Fax 1 250 need reschedule referred Dr. Name4 NamePattern1 NamePattern1 2539 Division Hematology evaluation anemia appointment scheduled 2 9 3 p.m. office located Location un Hospital Ward 23 Building Hospital1 18 Hospital Ward 516 Dr.[**Name NI 44536 administrative assistant Doctor 8982 Telephone/Fax 1 32192 need confirm reschedule

t-SNE visualization of the top 100 words from Word2Vec (MedSpaCy), with a limited word count for better label clarity.




```
#Build corpus
corpus_medspacy = build_corpus(patients_df_SciSpaCy, nlp_medspacy)
```

```
from gensim.models import Word2Vec
model_word2vec_medspacy = Word2Vec(corpus_medspacy, min_count=3)
```

```
len(model_word2vec_medspacy.wv.key_to_index.keys())
```

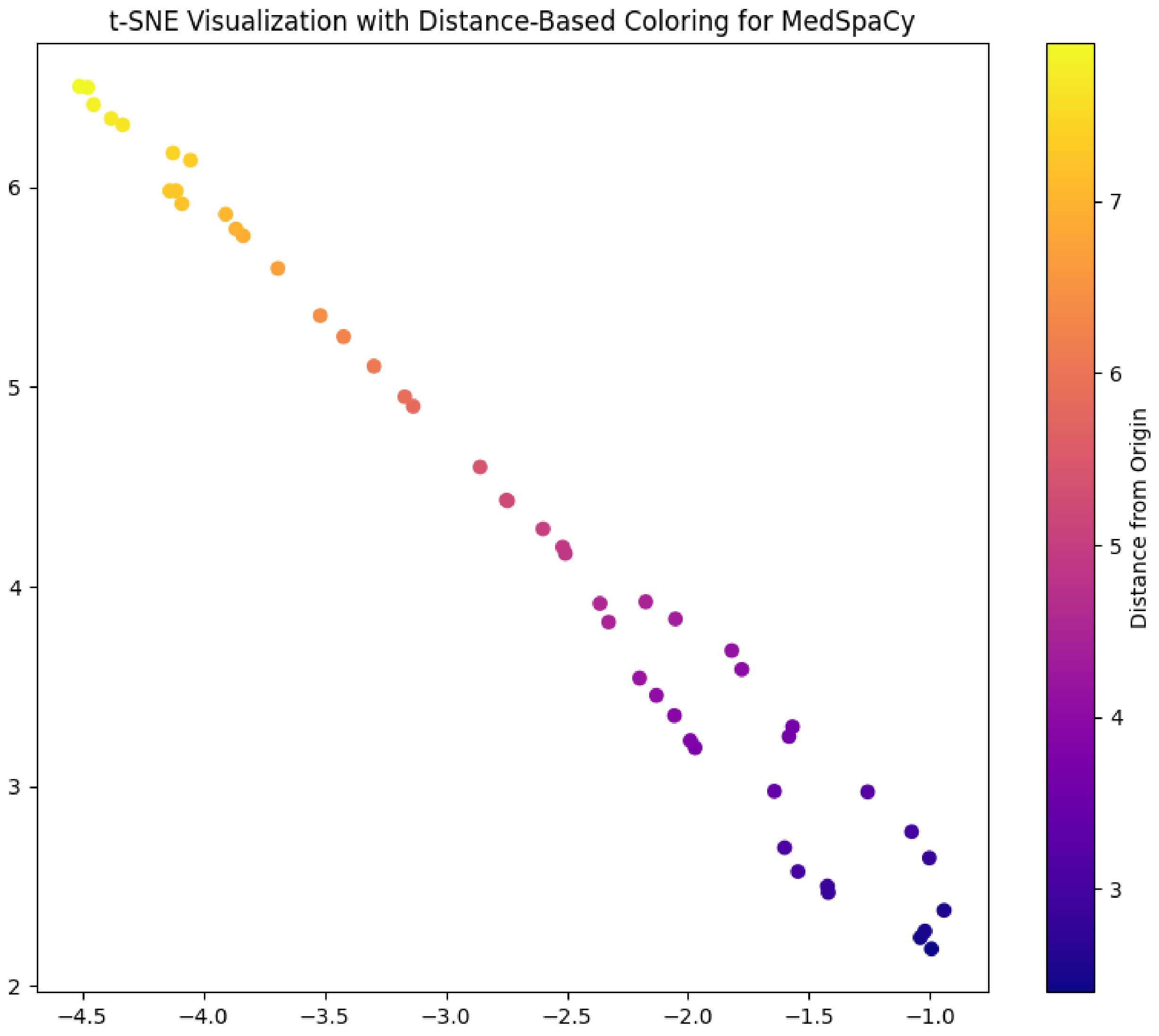
```
50
```

```
tsne_plot(model_word2vec_medspacy,np.array(list(model_word2vec_medspacy.wv.key_to_index.keys())), None, 'MedSpaCy')
```

Based on the plot, MedSpaCy effectively identifies and clusters terms associated with hypertension (e.g., Hypertension, Lisinopril, Labetalol, SBP, Clonidine, Headache, and Nausea). This suggests that the model is successfully recognizing and extracting a diverse range of hypertension-related entities, including medications, symptoms, and conditions, from the clinical text.

t-SNE Visualization with Distance-Based Coloring Of All Words from Word2Vec (MedSpaCy)

```
tsne_plot_no_label(model_word2vec_medspacy,np.array(list(model_word2vec_medspacy.wv.key_to_index.keys())), None, 'MedSpaCy')
```



BlueBert

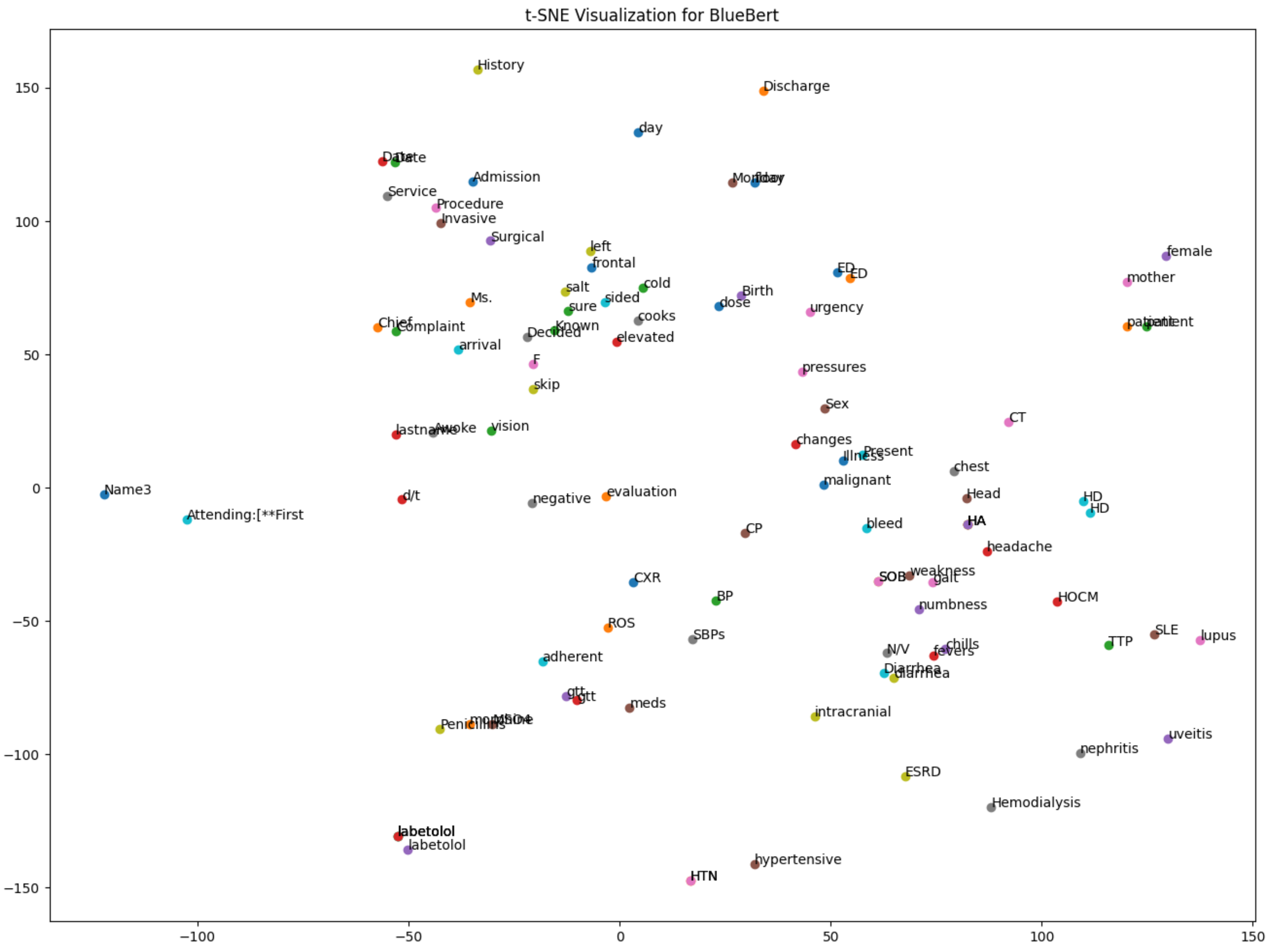
t-SNE Visualization For BlueBert

```python
# Visualization of notes filtered with SciSpacy using ClinicalBert
import numpy as np
import torch
from sklearn.manifold import TSNE
import string
import matplotlib.pyplot as plt
from transformers import AutoModel, AutoTokenizer, BertModel

# Load the BERT model and tokenizer
model_name = "bionlp/bluebert_pubmed_mimic_uncased_L-12_H-768_A-12"
tokenizer = AutoTokenizer.from_pretrained(model_name)
blue_bert_model = BertModel.from_pretrained('bionlp/bluebert_pubmed_mimic_uncased_L-12_H-768_A-12')
blue_bert_model.eval()

# Set first note as text
doc = nlp_SciSpaCy(patients_df_SciSpaCy['Processed_Text'][0])
corpus=[]
for ent in doc.ents:
    corpus.append(ent.text)
input_text =   ' '.join(corpus)

input_tokens = input_text.split()
word_embs = []

for token in input_tokens:
    # Check if the token is a valid word
    if token not in string.punctuation:
        # Encode the token using the BERT model
        inputs = tokenizer(token, return_tensors="pt")
        with torch.no_grad():
            outputs = blue_bert_model(**inputs)
        token_emb = outputs.last_hidden_state.mean(dim=1).squeeze().numpy()
        word_embs.append(token_emb)
```

- This script utilizes **BlueBERT** (bionlp/bluebert_pubmed_mimic_uncased_L-12_H-768_A-12) to extract word embeddings from clinical notes processed with SciSpaCy.
- Named entities are identified and tokenized, then their embeddings are computed using BlueBERT.
- **Only one note was used here because processing all notes with BlueBERT for embedding extraction requires significant time and memory.**
- The embeddings are visualized in a 2D space using t-SNE, highlighting relationships among clinical terms.

```python
# Perform t-SNE dimensionality reduction
tsne_model = TSNE(n_components=2, perplexity=10, random_state=42)
word_embs_2d = tsne_model.fit_transform(np.array(word_embs))
print(len(word_embs_2d))
# Create a scatter plot of the word embeddings in 2D space
plt.figure(figsize=(16,12))
for i in range(100):
    plt.scatter(word_embs_2d[i, 0], word_embs_2d[i, 1])
    plt.annotate(input_tokens[i], (word_embs_2d[i, 0], word_embs_2d[i, 1]))

plt.title(f"t-SNE Visualization for BlueBert")
plt.show()
```

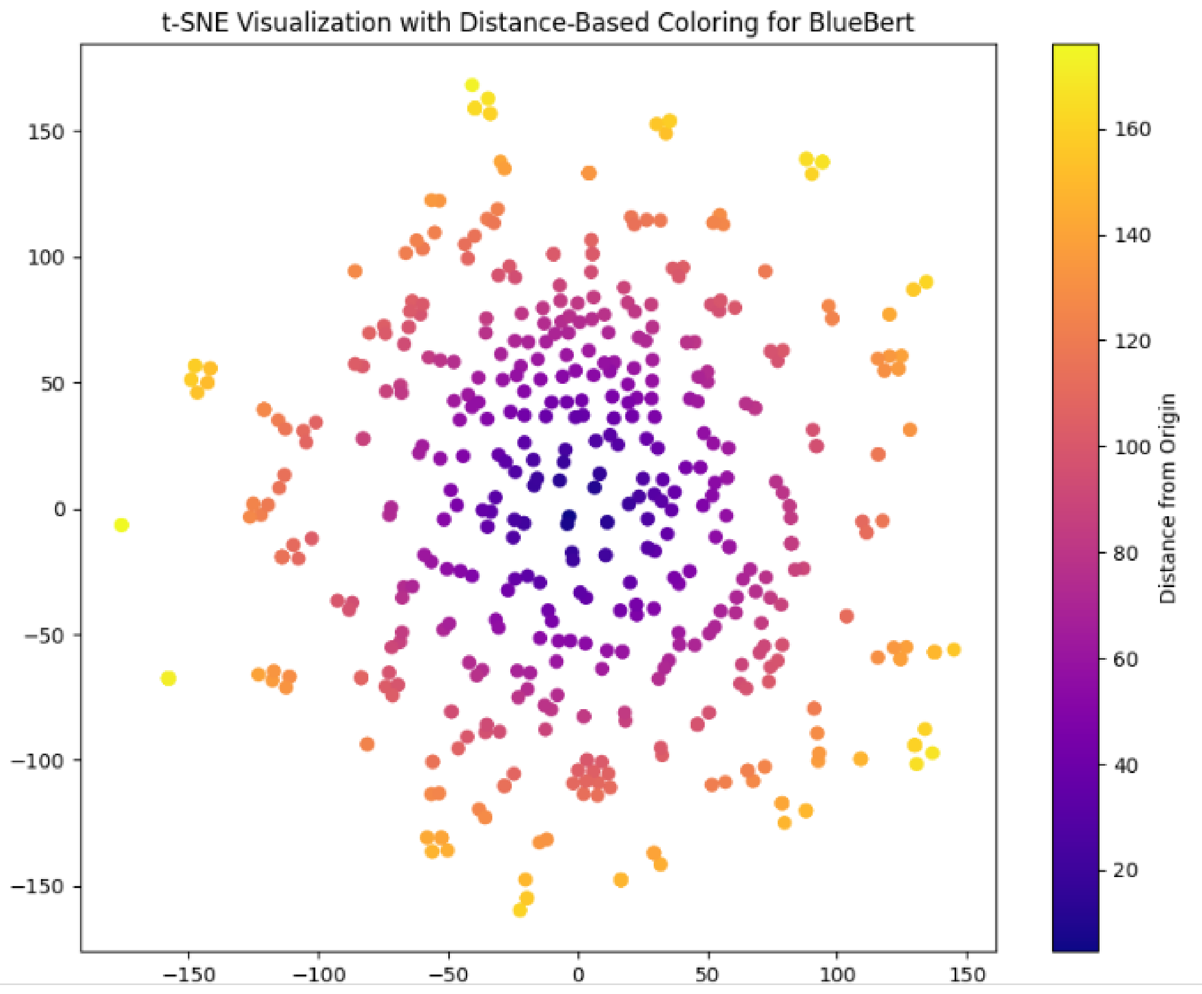
t-SNE visualization of the top 100 words from Word2Vec (BlueBert), with a limited word count for better label clarity.



Based on Word2Vec similarity and the plot, BlueBert effectively captures key medical terms related to hypertension, such as **Labetalol**, **hypertension**, and **hemodialysis**. It groups related terms like **HTN**, **hypertensive**, **BP**, **SBPs**, and medications like **Labetalol**, indicating its ability to identify and understand hypertension-related concepts.

t-SNE Visualization with Distance-Based Coloring Of All Words from Word2Vec (BlueBert)

```
plt.figure(figsize=(10, 8))
distances = np.sqrt(word_embs_2d[:, 0]**2 + word_embs_2d[:, 1]**2)
plt.scatter(word_embs_2d[:, 0], word_embs_2d[:, 1], c=distances, cmap='plasma')
plt.colorbar(label="Distance from Origin")
plt.title(f"t-SNE Visualization with Distance-Based Coloring for BlueBert")
plt.show()
```



Conclusion

The MIMIC data, especially the free-text notes, contains a lot of shorthand, misspellings, and extra details like dates and measurements that aren't useful for Named Entity Recognition (NER). Pre-trained models like BlueBERT, BC5CDR, and MedSpaCy, tailored for the medical field and charting terminology, tend to extract more relevant and accurate entities in NER than models like SpaCy and SciSpaCy.