AYDIN ADNAN MENDERES UNIVERSITY

ENGINEERING FACULTY
COMPUTER SCIENCE ENGINEERING DEPARTMENT

NER On Medical Text

CSE431 – Natural Language Processing with Machine Learning 2023/2024

Burak TÜZEL
Talha Alper ASAV

Lecturer:

Asst. Prof. Dr. Fatih SOYGAZİ

Named Entity Recognition on Medical Text

Installing necessary environment and importing the libraries:

1- Install Jupyter Notebook

```
:  import spacy
   import scispacy
   #Core models
   import en_core_sci_sm
   import en_core_sci_md
   #NER specific models
   import en_ner_bc5cdr_md
   #Tools for extracting & displaying data
   from spacy import displacy
   import pandas as pd
```

Downloading the Medical Text:

Here is the medical text link to download: https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions

```
# Reading the csv file
file_path = 'D:/nlp/mtsamples.csv'
df = pd.read_csv(file_path)
```

```
print(df.head())
```

```
   Unnamed: 0                                        description  \
0           0   A 23-year-old white female presents with comp...
1           1           Consult for laparoscopic gastric bypass.
2           2           Consult for laparoscopic gastric bypass.
3           3                             2-D M-Mode. Doppler.
4           4                                 2-D Echocardiogram

            medical_specialty                                sample_name  \
0        Allergy / Immunology                          Allergic Rhinitis
1                  Bariatrics   Laparoscopic Gastric Bypass Consult - 2
2                  Bariatrics   Laparoscopic Gastric Bypass Consult - 1
3   Cardiovascular / Pulmonary                    2-D Echocardiogram - 1
4   Cardiovascular / Pulmonary                    2-D Echocardiogram - 2

                                      transcription  \
0  SUBJECTIVE:,  This 23-year-old white female pr...
1  PAST MEDICAL HISTORY:, He has difficulty climb...
2  HISTORY OF PRESENT ILLNESS: , I have seen ABC ...
3  2-D M-MODE: , ,1.  Left atrial enlargement wit...
4  1.  The left ventricular cavity size and wall ...

                                        keywords
0  allergy / immunology, allergic rhinitis, aller...
1  bariatrics, laparoscopic gastric bypass, weigh...
2  bariatrics, laparoscopic gastric bypass, heart...
3  cardiovascular / pulmonary, 2-d m-mode, dopple...
4  cardiovascular / pulmonary, 2-d, doppler, echo...
```

Finding the Disease, Drugs and Drugs-Doses Named Entities:

This one for “en_core_sci_sm” model:

```
nlp_sm = en_core_sci_sm.load()
doc = nlp_sm(text)
displacy_image = displacy.render(doc, jupyter=True,style='ent')
```

2-D ENTITY M-MODE: , ,1. Left atrial enlargement ENTITY with left atrial diameter ENTITY of 4.7 cm.,2 ENTITY . Normal size right ENTITY and left ventricle.,3 ENTITY . Normal LV systolic function with left ventricular ejection fraction ENTITY of 51%.,4. Normal LV diastolic function.,5. No pericardial effusion.,6. Normal morphology ENTITY of aortic valve ENTITY , mitral valve ENTITY , tricuspid valve ENTITY , and pulmonary valve.,7 ENTITY . PA systolic pressure is 36 mmHg.,DOPPLER: , ,1. Mild mitral ENTITY and tricuspid regurgitation.,2 ENTITY . Trace aortic ENTITY and pulmonary regurgitation ENTITY .

This one for “en_core_sci_md” model:

```
nlp_md = en_core_sci_md.load()
doc = nlp_md(text)
#Display resulting entity extraction
displacy_image = displacy.render(doc, jupyter=True,style='ent')
```

2-D ENTITY M-MODE: , ,1. Left atrial enlargement ENTITY with left atrial ENTITY diameter ENTITY of 4.7 cm.,2 ENTITY . Normal size right ENTITY and left ventricle.,3 ENTITY . Normal LV systolic function with left ventricular ejection fraction ENTITY of 51%.,4 ENTITY . Normal LV diastolic function.,5. No pericardial effusion.,6. Normal morphology ENTITY of aortic valve ENTITY , mitral valve ENTITY , tricuspid valve ENTITY , and pulmonary valve.,7 ENTITY . PA systolic pressure is 36 mmHg.,DOPPLER ENTITY : , ,1. Mild mitral and tricuspid regurgitation.,2 ENTITY . Trace aortic and pulmonary regurgitation ENTITY .

This one for “en_ner_bc5cdr_md” model:

```
nlp_bc = en_ner_bc5cdr_md.load()
doc = nlp_bc(text)
#Display resulting entity extraction
displacy_image = displacy.render(doc, jupyter=True,style='ent')
```

2-D M-MODE: , ,1. Left atrial enlargement DISEASE with left atrial diameter of 4.7 cm.,2. Normal size right and left ventricle.,3. Normal LV systolic function with left ventricular ejection fraction of 51%.,4. Normal LV diastolic function.,5. No pericardial effusion.,6. Normal morphology of aortic valve, mitral valve, tricuspid valve, and pulmonary valve.,7. PA systolic pressure is 36 mmHg.,DOPPLER: , ,1. Mild mitral and tricuspid regurgitation.,2. Trace aortic and pulmonary regurgitation DISEASE .

Here is the Result:

```
df.dropna(subset=['transcription'], inplace=True)
df_subset = df.sample(n=100, replace=False, random_state=42)
df_subset.info()
df_subset.head()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 100 entries, 3162 to 3581
Data columns (total 6 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Unnamed: 0         100 non-null    int64 
 1   description        100 non-null    object
 2   medical_specialty  100 non-null    object
 3   sample_name        100 non-null    object
 4   transcription      100 non-null    object
 5   keywords           78 non-null     object
dtypes: int64(1), object(5)
memory usage: 5.5+ KB
```

	Unnamed: 0	description	medical_specialty	sample_name	transcription	keywords
3162	3162	Markedly elevated PT INR despite stopping Cou...	Hematology - Oncology	Hematology Consult - 1	HISTORY OF PRESENT ILLNESS:, The patient is w...	NaN
1981	1981	Intercostal block from fourth to tenth interc...	Pain Management	Intercostal block - 1	PREPROCEDURE DIAGNOSIS:, Chest pain secondary...	pain management, xylocaine, marcaine, intercos...
1361	1361	The patient is a 65-year-old female who under...	SOAP / Chart / Progress Notes	Lobectomy - Followup	HISTORY OF PRESENT ILLNESS: , The patient is a...	soap / chart / progress notes, non-small cell ...
3008	3008	Construction of right upper arm hemodialysis ...	Nephrology	Hemodialysis Fistula Construction	PREOPERATIVE DIAGNOSIS: , End-stage renal dise...	nephrology, end-stage renal disease, av dialys...
4943	4943	Bronchoscopy with brush biopsies. Persistent...	Cardiovascular / Pulmonary	Bronchoscopy - 8	PREOPERATIVE DIAGNOSIS: , Persistent pneumonia...	cardiovascular / pulmonary, persistent pneumon...

```
from spacy.matcher import Matcher
pattern = [{'ENT_TYPE':'CHEMICAL'}, {'LIKE_NUM': True}, {'IS_ASCII': True}]
matcher = Matcher(nlp_bc.vocab)
matcher.add("DRUG_DOSE", [pattern])
for transcription in df_subset['transcription']:
    doc = nlp_bc(transcription)
    matches = matcher(doc)
    for match_id, start, end in matches:
        string_id = nlp_bc.vocab.strings[match_id]  # get string representation
        span = doc[start:end]  # the matched span adding drugs doses
        print(span.text, start, end, string_id,)
        #Add disease and drugs
        for ent in doc.ents:
            print(ent.text, ent.start_char, ent.end_char, ent.label_)
```

```
Xylocaine 20 mL 129 132 DRUG_DOSE
Chest pain 26 36 DISEASE
Chest pain 122 132 DISEASE
intercostal block 318 335 DISEASE
chest pain 388 398 DISEASE
Xylocaine 730 739 CHEMICAL
Marcaine 750 758 CHEMICAL
contusion 987 996 DISEASE
respiratory distress 1076 1096 DISEASE
pain 1150 1154 DISEASE
Marcaine 0.25% 133 136 DRUG_DOSE
Chest pain 26 36 DISEASE
Chest pain 122 132 DISEASE
intercostal block 318 335 DISEASE
chest pain 388 398 DISEASE
Xylocaine 730 739 CHEMICAL
Marcaine 750 758 CHEMICAL
contusion 987 996 DISEASE
```

Then We Saved the Output as .csv File So We Can Use Later On:

```
import csv
from spacy.matcher import Matcher

# Your existing code for creating matcher and processing text
pattern = [{'ENT_TYPE': 'CHEMICAL'}, {'LIKE_NUM': True}, {'IS_ASCII': True}]
matcher = Matcher(nlp_bc.vocab)
matcher.add("DRUG_DOSE", [pattern])

# Open a CSV file for writing
csv_file_path = 'output.csv'
with open(csv_file_path, 'w', newline='') as csv_file:
    csv_writer = csv.writer(csv_file)

    # Write header row
    csv_writer.writerow(['Text', 'Start', 'End', 'Entity Type'])

    for transcription in df_subset['transcription']:
        doc = nlp_bc(transcription)
        matches = matcher(doc)

        for match_id, start, end in matches:
            string_id = nlp_bc.vocab.strings[match_id]
            span = doc[start:end]

            # Write to CSV
            csv_writer.writerow([span.text, start, end, string_id])

            # Add disease and drugs information to the CSV
            for ent in doc.ents:
                csv_writer.writerow([ent.text, ent.start_char, ent.end_char, ent.label_])
```

REFERENCES

1. https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions
2. https://www.analyticsvidhya.com/blog/2023/02/extracting-medical-information-from-clinical-text-with-nlp/