FR. CONCEICAO RODRIGUES COLLEGE OF ENGINEERIG
Department of Computer Engineering

Experiment 5 and 6 - Natural Language Entity Extraction from Medical Reports

Course Details:

Academic Year	**2023 - 24**	**Estimated Time**	**Experiment No. 5 & 6 – 02 Hours**
Course & Semester	**B.E. (COMP) – Sem. VII**	**Subject Name**	**Data Science for Health and Social Care Lab**
Experiment Type	**Software Performance**	**Subject Code**	**HDSSBL701**

Name of Student	Atharva Pawar	**Roll No.**	9427
Date of Performance.:		**Date of Submission.:**	
CO Mapping	HDSSBL701.2 Clean, integrate and transform healthcare data. HDSSBL701.5 Implement data science solutions for solving healthcare problems.		

Aim: Natural Language Entity Extraction from Medical Reports

Objective: To perform the process of entity extraction from medical text data using NLP techniques and tools..

Tools and Libraries:

- Medical text data (e.g., sample medical reports or synthetic data)
- NLP libraries (e.g., spaCy, NLTK, or Hugging Face Transformers)
- Pre-trained NLP models (e.g., spaCy's "en_core_med7" or BERT-based models)
- Evaluation metrics (precision, recall, F1-score)

Procedure:

Step 1: Data Preparation

Collect a dataset of medical reports in text format.

Understand the structure of the data, including the types of entities needed to extract (e.g., medical conditions, medications, dates).

Step 2: Data Cleaning

Remove any irrelevant information such as headers, footers, page numbers, and boilerplate text.

Remove special characters, symbols, and non-alphanumeric characters that don't contribute to the analysis.

Handle or remove noisy text like HTML tags, XML markup, or other formatting artifacts.

Step 3: Text Pre-processing

a. Tokenization: Split the cleaned text into individual words or tokens.

Consider using more specialized tokenization methods for medical terms or abbreviations.

b. Lowercasing: Convert all text to lowercase to ensure consistent processing.

Stopword Removal:

c. Remove common stopwords that don't contribute much to the overall meaning of the text. Create a custom stopword list that considers medical domain-specific terms.
d. Stemming and Lemmatization: Apply stemming or lemmatization to reduce words to their root forms.Consider using a medical-specific stemmer or lemmatizer if available.
e. Spell Checking and Correction: Implement spell checking and correction techniques to fix common typos or misspellings. Utilize medical dictionaries to ensure accurate corrections for domain-specific terms.
f. Handling Numeric Data:Identify and handle numeric values such as measurements, lab values, and vital signs. Normalize numeric values to a consistent format.
g. Handling Dates and Times: Extract and standardize date and time information from the text. Convert dates to a common format for analysis.
h. Removing Personal Identifiers: Anonymize or remove personally identifiable information (PII) to ensure privacy and compliance with data protection regulations.
i. Handling Missing Data: Decide on strategies for dealing with missing data, such as filling with placeholders or removing affected records.

Step 4: Entity Recognition and Normalization

a. Identify and label medical entities like diseases, treatments, medications, and anatomical terms. Use Named Entity Recognition (NER) tools using libraries like spaCy or specialized NER tools, or libraries trained on medical data if available.
b. Standardize abbreviations, acronyms, and variations of medical terms. Map synonyms to a common representation.

Step 5: Visualizing the impact of data cleaning and pre-processing.

Step 6: Evaluation:

Evaluate the performance of the entity extraction system. Use precision, recall, and F1-score for their extracted entities. Fine-tune the models to improve results.

Dataset:
https://drive.google.com/file/d/1VMu8eLNLfk1SQUJ8HldFYYiEX_WYWKW7/view?usp=drive_link

Reference Links for Implementation:
https://www.analyticsvidhya.com/blog/2019/03/learn-to-use-elmo-to-extract-features-from-text/
https://www.analyticsvidhya.com/blog/2023/02/extracting-medical-information-from-clinical-text-with-nlp/

DSHC - exp - 5 and 6

In [66]:
```
import pandas as pd

# Load the CSV file into a DataFrame
df = pd.read_csv('mtsamples.csv')

# Select only the "transcription" column
trans_col = df['transcription']
```

In [67]:
```
df.head
```

Out[67]:
```
<bound method NDFrame.head of       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
...          ...                                                ...
4994        4994   Patient having severe sinusitis about two to ...
4995        4995   This is a 14-month-old baby boy Caucasian who...
4996        4996   A female for a complete physical and follow u...
4997        4997   Mother states he has been wheezing and coughing.
4998        4998   Acute allergic reaction, etiology uncertain, ...

                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
...                           ...                                       ...
4994        Allergy / Immunology                          Chronic Sinusitis
4995        Allergy / Immunology      Kawasaki Disease - Discharge Summary
4996        Allergy / Immunology                          Followup on Asthma
4997        Allergy / Immunology                       Asthma in a 5-year-old
4998        Allergy / Immunology                Allergy Evaluation Consult

                                          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 ...
...                                                 ...
4994  HISTORY:,  I had the pleasure of meeting and e...
4995  ADMITTING DIAGNOSIS: , Kawasaki disease.,DISCH...
4996  SUBJECTIVE: , This is a 42-year-old white fema...
4997  CHIEF COMPLAINT: , This 5-year-old male presen...
4998  HISTORY: , A 34-year-old male presents today s...

                                             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...
...                                                 ...
4994                                              NaN
4995  allergy / immunology, mucous membranes, conjun...
4996                                              NaN
4997                                              NaN
4998                                              NaN

[4999 rows x 6 columns]>
```

In [68]:
```
df.tail
```

Out[68]:
```
<bound method NDFrame.tail of       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
...          ...                                                ...
4994        4994   Patient having severe sinusitis about two to ...
4995        4995   This is a 14-month-old baby boy Caucasian who...
4996        4996   A female for a complete physical and follow u...
4997        4997   Mother states he has been wheezing and coughing.
4998        4998   Acute allergic reaction, etiology uncertain, ...

                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
...                          ...                                    ...
4994         Allergy / Immunology                       Chronic Sinusitis
4995         Allergy / Immunology     Kawasaki Disease - Discharge Summary
4996         Allergy / Immunology                     Followup on Asthma
4997         Allergy / Immunology                  Asthma in a 5-year-old
4998         Allergy / Immunology              Allergy Evaluation Consult

                                        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 ...
...                                                 ...
4994  HISTORY:,  I had the pleasure of meeting and e...
4995  ADMITTING DIAGNOSIS: , Kawasaki disease.,DISCH...
4996  SUBJECTIVE: , This is a 42-year-old white fema...
4997  CHIEF COMPLAINT: , This 5-year-old male presen...
4998  HISTORY: , A 34-year-old male presents today s...

                                              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...
...                                                 ...
4994                                              NaN
4995  allergy / immunology, mucous membranes, conjun...
4996                                              NaN
4997                                              NaN
4998                                              NaN

[4999 rows x 6 columns]>
```

In [69]:
```
# Display the first 5 rows
print(trans_col.head(10))
```

```
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 ...
5    PREOPERATIVE DIAGNOSIS: , Morbid obesity.,POST...
6    PREOPERATIVE DIAGNOSES:,1.  Deformity, right b...
7    2-D ECHOCARDIOGRAM,Multiple views of the heart...
8    PREOPERATIVE DIAGNOSIS: , Lipodystrophy of the...
9    DESCRIPTION:,1.  Normal cardiac chambers size....
Name: transcription, dtype: object
```

In [70]:
```
trans_col[0]
```

Out[70]:
```
'SUBJECTIVE:,  This 23-year-old white female presents with complaint of allergies.  She used to have allergies when she lived i
n Seattle but she thinks they are worse here.  In the past, she has tried Claritin, and Zyrtec.  Both worked for short time but
then seemed to lose effectiveness.  She has used Allegra also.  She used that last summer and she began using it again two week
s ago.  It does not appear to be working very well.  She has used over-the-counter sprays but no prescription nasal sprays.  Sh
e does have asthma but doest not require daily medication for this and does not think it is flaring up.,MEDICATIONS: , Her only
medication currently is Ortho Tri-Cyclen and the Allegra.,ALLERGIES: , She has no known medicine allergies.,OBJECTIVE:,Vitals:
Weight was 130 pounds and blood pressure 124/78.,HEENT:  Her throat was mildly erythematous without exudate.  Nasal mucosa was
erythematous and swollen.  Only clear drainage was seen.  TMs were clear.,Neck:  Supple without adenopathy.,Lungs:  Clear.,ASSE
SSMENT:,  Allergic rhinitis.,PLAN:,1.  She will try Zyrtec instead of Allegra again.  Another option will be to use loratadine.
She does not think she has prescription coverage so that might be cheaper.,2.  Samples of Nasonex two sprays in each nostril gi
ven for three weeks.  A prescription was written as well.'
```

In [71]:
```
# Display the last 5 rows
print(trans_col.tail(5))
```

```
4994    HISTORY:,  I had the pleasure of meeting and e...
4995    ADMITTING DIAGNOSIS: , Kawasaki disease.,DISCH...
4996    SUBJECTIVE: , This is a 42-year-old white fema...
4997    CHIEF COMPLAINT: , This 5-year-old male presen...
4998    HISTORY: , A 34-year-old male presents today s...
Name: transcription, dtype: object
```

In [72]:
```
trans_col.describe()
```

Out[72]:
```
count                                          4966
unique                                         2357
top       PREOPERATIVE DIAGNOSIS: , Low back pain.,POSTO...
freq                                              5
Name: transcription, dtype: object
```

In [73]:
```
import nltk

# Download the NLTK tokenizer model (if not already downloaded)
nltk.download('punkt')

# Check for NaN values and replace them with empty strings
trans_col = trans_col.fillna('')

# Tokenize the "transcription" column
```

```
trans_col_tokenize = trans_col.apply(lambda x: nltk.word_tokenize(str(x)))
```

```
[nltk_data] Downloading package punkt to C:\Users\Atharva
[nltk_data]     Pawar\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

In [74]:
```
trans_col_tokenize[0]
```

Out[74]:
```
['SUBJECTIVE',
 ':',
 ',',
 'This',
 '23-year-old',
 'white',
 'female',
 'presents',
 'with',
 'complaint',
 'of',
 'allergies',
 '.',
 'She',
 'used',
 'to',
 'have',
 'allergies',
 'when',
```

In [75]:
```
# Convert text to lowercase in the "transcription" column
trans_col_tok_lowC = trans_col_tokenize.apply(lambda x: [word.lower() for word in x])
trans_col_tok_lowC[0]
```

Out[75]:
```
['subjective',
 ':',
 ',',
 'this',
 '23-year-old',
 'white',
 'female',
 'presents',
 'with',
 'complaint',
 'of',
 'allergies',
 '.',
 'she',
 'used',
 'to',
 'have',
 'allergies',
 'when',
```

In [76]:
```
# Download the NLTK stopwords corpus (if not already downloaded)
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to C:\Users\Atharva
[nltk_data]     Pawar\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[76]:
```
True
```

In [77]:
```
# Import the NLTK stopwords
from nltk.corpus import stopwords

# Define a custom list of medical domain-specific stopwords
custom_stopwords = ["medical_term1", "medical_term2", "medical_term3", ...]

# Get the standard English stopwords
english_stopwords = set(stopwords.words('english'))

# Combine the custom medical stopwords and the standard English stopwords
all_stopwords = set(custom_stopwords).union(english_stopwords)

# Remove stopwords from the "transcription" column using the custom list
trans_col_tok_lowC_stopW = trans_col_tok_lowC.apply(lambda x: [word for word in x if word not in all_stopwords])
trans_col_tok_lowC_stopW[0]
```

Out[77]:
```
['subjective',
 ':',
 ',',
 '23-year-old',
 'white',
 'female',
 'presents',
 'complaint',
 'allergies',
 '.',
 'used',
 'allergies',
 'lived',
 'seattle',
 'thinks',
 'worse',
 '.',
```

```
 'past',
 ',',
```

In [78]:
```
import spacy

# Load the spaCy English language model
nlp = spacy.load("en_core_web_sm")
```

In [79]:
```
# Define a function for lemmatization
def lemmatize_text(text):
    doc = nlp(" ".join(text))
    return [token.lemma_ for token in doc]

# Apply lemmatization to the "transcription" column
trans_col_tok_lowC_stopW_Lemm = trans_col_tok_lowC_stopW.apply(lemmatize_text)
trans_col_tok_lowC_stopW_Lemm[0]
```

Out[79]:
```
['subjective',
 ':',
 ',',
 '23',
 '-',
 'year',
 '-',
 'old',
 'white',
 'female',
 'present',
 'complaint',
 'allergy',
 '.',
 'use',
 'allergy',
 'live',
 'seattle',
 'think',
```

In [80]:
```
# !pip install indexer
# !pip install pyspellchecker
```

In [81]:
```
# import re
from spellchecker import SpellChecker
# import dateutil.parser
```

In [82]:
```
# Spell Checking and Correction
def spell_check_and_correct(text):
    spell = SpellChecker()
    words = text.split()
    corrected_words = [spell.correction(word) if word is not None else word for word in words]
    corrected_words = [word for word in corrected_words if word is not None]  # Filter out None values
    corrected_text = ' '.join(corrected_words)
    return corrected_text
```

In [83]:
```
# Handling Numeric Data
def extract_numeric_data(text):
    numeric_values = re.findall(r'\d+\.\d+|\d+', text)
    return [float(value) for value in numeric_values]
```

In [84]:
```
# Handling Dates and Times
def extract_dates_and_times(text):
    try:
        dates = dateutil.parser.parse(text, fuzzy=True)
        return dates
    except:
        return None
```

In [85]:
```
# Removing Personal Identifiers (Names in this example)
def remove_names(text):
    name_pattern = re.compile(r'\b[A-Z][a-z]+\b')
    return name_pattern.sub('XXXX', text)
```

In [86]:
```
# Handling Missing Data (Fill with Placeholder in this example)
def fill_missing_data(text, placeholder="[MISSING]"):
    return text.replace("MISSING", placeholder)
```

In [87]:
```
# Sample Text
sample_text = "A 23-year-old white female presents with complaint of allergies. Allergy / Immunology. Allergic Rhinitis. SUBJECTI

# Apply Functions
sample_text = spell_check_and_correct(sample_text)
numeric_data = extract_numeric_data(sample_text)
dates_and_times = extract_dates_and_times(sample_text)
sample_text = remove_names(sample_text)
sample_text = fill_missing_data(sample_text)

# Output
print("Spell Checked and Corrected Text:")
print(sample_text)
print("\nExtracted Numeric Data:")
print(numeric_data)
```

```
print("\nExtracted Dates and Times: ")
print(dates_and_times)
# This script defines the functions and demonstrates their use with a sample text. You can adapt and use these functions accordir
```

```
Spell Checked and Corrected Text:
A white female presents with complaint of allergies XXXX / immunology XXXX subjective XXXX white female presents with complaint
of allergies XXXX used to have allergies when she lived in XXXX but she thinks they are worse here XXXX has tried clarity and X
XXX worked for a short time but then seemed to lose effectiveness XXXX has used XXXX also XXXX used that last summer and she be
gan using it again two weeks ago XXXX does not appear to be working very well XXXX has used sprays but no prescription nasal sp
rays XXXX does have asthma but does not require daily medication for this and does not think it is flaring up

Extracted Numeric Data:
[]

Extracted Dates and Times:
None
```

In [88]:
```
# Download the NER corpus if not already downloaded
nltk.download('maxent_ne_chunker')
nltk.download('words')
```

```
[nltk_data] Downloading package maxent_ne_chunker to C:\Users\Atharva
[nltk_data]     Pawar\AppData\Roaming\nltk_data...
[nltk_data]   Package maxent_ne_chunker is already up-to-date!
[nltk_data] Downloading package words to C:\Users\Atharva
[nltk_data]     Pawar\AppData\Roaming\nltk_data...
[nltk_data]   Package words is already up-to-date!
```

Out[88]:
```
True
```

In [89]:
```
import nltk


from nltk import word_tokenize, pos_tag, ne_chunk

def custom_ner(text):
    words = word_tokenize(text)
    tagged = pos_tag(words)
    named_entities = ne_chunk(tagged)

    return named_entities

# # Sample text for NER
# sample_text = "Barack Obama was born in Honolulu. He was the President of the United States."

# # Apply the custom NER function
# entities = custom_ner(sample_text)

# # Print the named entities
# print(entities)
```

In [90]:
```
# Apply the custom NER function to the 'description' column
df['NER'] = df['description'].apply(custom_ner)

print(df['NER'])
```

```
0       [(A, DT), (23-year-old, JJ), (white, JJ), (fem...
1       [[(Consult, NN)], (for, IN), (laparoscopic, NN...
2       [[(Consult, NN)], (for, IN), (laparoscopic, NN...
3       [(2-D, JJ), (M-Mode, NNP), (., .), [(Doppler, ...
4                       [(2-D, JJ), (Echocardiogram, NNP)]
                              ...
4994    [(Patient, NNP), (having, VBG), (severe, JJ), ...
4995    [(This, DT), (is, VBZ), (a, DT), (14-month-old...
4996    [(A, DT), (female, NN), (for, IN), (a, DT), (c...
4997    [(Mother, RB), (states, NNS), (he, PRP), (has,...
4998    [[(Acute, NNP)], (allergic, JJ), (reaction, NN...
Name: NER, Length: 4999, dtype: object
```

In [91]:
```
ner_row_2 = df.loc[1, 'NER']  # Access row 2 (index 1) and the 'NER' column
print(ner_row_2)
```

```
(S (GSP Consult/NN) for/IN laparoscopic/NN gastric/JJ bypass/NN ./.)
```

In [ ]:

In [ ]: