🔬 Text Classification using N-Gram Feature Extraction 📝

Summary

Medical domain is in a data rich environment that a variety of knowledge can be extracted for positive outcomes. This notebook work will show multiclass classification of medical transcriptions using a real dataset. The objective of this paper is to classify medical transcriptions based on the medical specialty labels, namely Discharge Summary, Neurosurgery and ENT. Text normalisation has performed followed by extracting five different n-gram feature representations are. Moreover, three supervised learning classifiers were trained on each of the n-gram feature representations, namely K-Nearest Neighbours, Decision Tree, and Random Forest. The classification performance was evaluated by the metric score of macro F1. The best score achieved was over 0.8 macro F1 on testing set using tuned Random Forest and unigram feature vectors.

Concept of N-Gram

In this notebook, n-gram will be used for feature extraction, which is the first NLP approach that introduced by Markov in 1913 [1]. An N-gram is an N-character slice of a longer string. The intuition of the n-gram model is that instead of computing a prediction based on entire corpus, one can approximate the prediction by only contiguous slices sequence of n words [2]. To explain feature extraction using n-gram with a demonstration of the sentence, “The student is alone happily”. The number of n-gram features can be calculated by k-n+1, where k is the number of words. The result is a bag-of-n-grams model [3] for a classifier to train the linguistic algorithm. Table I below shows the demonstration of different n-gram feature representations.

Contents


```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
import pandas as pd
```

```
import numpy as np
import re
import warnings
import matplotlib
import matplotlib.pyplot as plt
from nltk.tokenize import WhitespaceTokenizer
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
from wordcloud import WordCloud
from sklearn import preprocessing
from sklearn.feature_extraction.text import CountVectorizer
import os

def trim(df):
    df.columns = df.columns.str.strip()
    df = df.drop_duplicates()
    df.columns = df.columns.str.lower()
    df.columns = df.columns.str.replace(' ','_')
    df_obj = df.select_dtypes(['object'])
    df[df_obj.columns] = df_obj.apply(lambda x: x.str.strip())
    print("All column names have been striped, lowered case, replaced space with underscore if any")
    print("Dropped duplicated instances if any")
    print("Categorical instances have been striped")
    return df

pd.set_option('display.max_colwidth', 255)
df =pd.read_csv('drive/MyDrive/mtsamples.csv')
df.drop('Unnamed: 0', axis=1, inplace=True)
df = trim(df)

def vc(df, column, r=False):
    vc_df = df.reset_index().groupby([column]).size().to_frame('count')
    vc_df['percentage (%)'] = vc_df['count'].div(sum(vc_df['count'])).mul(100)
    vc_df = vc_df.sort_values(by=['percentage (%)'], ascending=False)
    if r:
        return vc_df
    else:
        print(f'STATUS: Value counts of "{column}"...')
        display(vc_df)

def shape(df,df_name):
    print(f'STATUS: Dimension of "{df_name}" = {df.shape}')

df.head(1000)
```

```
All column names have been striped, lowered case, replaced space with underscore if any
Dropped duplicated instances if any
Categorical instances have been striped
```

	description	medical_specialty	sample_name	transcription	keywords
0	A 23-year-old white female presents with complaint of allergies.	Allergy / Immunology	Allergic Rhinitis	SUBJECTIVE:, This 23-year-old white female presents with complaint of allergies. She used to have allergies when she lived in Seattle but she thinks they are worse here. In the past, she has tried Claritin, and Zyrtec. Both worked for short time b...	allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal, erythematous, allegra, sprays, allergic,
1	Consult for laparoscopic gastric bypass.	Bariatrics	Laparoscopic Gastric Bypass Consult - 2	PAST MEDICAL HISTORY:, He has difficulty climbing stairs, difficulty with airline seats, tying shoes, used to public seating, and lifting objects off the floor. He exercises three times a week at home and does cardio. He has difficulty walking two b...	bariatrics, laparoscopic gastric bypass, weight loss programs, gastric bypass, atkin's diet, weight watcher's, body weight, laparoscopic gastric, weight loss, pounds, months, weight, laparoscopic, band, loss, diets, overweight, lost
2	Consult for laparoscopic gastric bypass.	Bariatrics	Laparoscopic Gastric Bypass Consult - 1	HISTORY OF PRESENT ILLNESS: , I have seen ABC today. He is a very pleasant gentleman who is 42 years old, 344 pounds. He is 5'9". He has a BMI of 51. He has been overweight for ten years since the age of 33, at his highest he was 358 pounds, at hi...	bariatrics, laparoscopic gastric bypass, heart attacks, body weight, pulmonary embolism, potential complications, sleep study, weight loss, gastric bypass, anastomosis, loss, sleep, laparoscopic, gastric, bypass, heart, pounds, weight,
3	2-D M-Mode. Doppler.	Cardiovascular / Pulmonary	2-D Echocardiogram - 1	2-D M-MODE: , ,1. Left atrial enlargement 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 pericard...	cardiovascular / pulmonary, 2-d m-mode, doppler, aortic valve, atrial enlargement, diastolic function, ejection fraction, mitral, mitral valve, pericardial effusion, pulmonary valve, regurgitation, systolic function, tricuspid, tricuspid valve, normal lv
4	2-D Echocardiogram	Cardiovascular / Pulmonary	2-D Echocardiogram - 2	1. The left ventricular cavity size and wall thickness appear normal. The wall motion and left ventricular systolic function appears hyperdynamic with estimated ejection fraction of 70% to 75%. There is near-cavity obliteration seen. There also ap...	cardiovascular / pulmonary, 2-d, doppler, echocardiogram, annular, aortic root, aortic valve, atrial, atrium, calcification, cavity, ejection fraction, mitral, obliteration, outflow, regurgitation, relaxation pattern, stenosis, systolic function, tric...
...	...	...	...	...	...
					surgery, sulfasalazine

Patient with active flare of Inflammatory Bowel

PROCEDURES PERFORMED:, Colonoscopy.,INDICATIONS:,

cortisone local therapy, inflammatory bowel disease,

```
df = df[df['medical_specialty'].isin(['Neurosurgery','ENT - Otolaryngology','Discharge Summary'])]
shape(df,'df')
```

```
     STATUS: Dimension of "df" = (300, 5)
```

inflammatory, rectal,

1.2 | Word Counts of Each Medical Specialty

surgery, diarrhea,

To query the data, I would like to know how is the size of the dataset and also to rank null values in descending order

Diarrhea, suspected irritable colonoscopy,

```
medical_specialty_list = [] ; word_count_list =[]
for medical_specialty in df['medical_specialty'].unique():
    df_filter = df.loc[(df['medical_specialty'] == medical_specialty)]
    word_count_temp = df_filter['transcription'].str.split().str.len().sum()
    medical_specialty_list.append(medical_specialty)
    word_count_list.append(word_count_temp)
word_count_df = pd.DataFrame({'Medical Specialty':medical_specialty_list, 'Word Count':word_count_list})
word_count_df['Word Count'] = word_count_df['Word Count'].astype('int')
word_count_df = word_count_df.sort_values('Word Count', ascending=False)
word_count_df.reset_index(drop=True)
```

	Medical Specialty	Word Count
0	Surgery	526754
1	Consult - History and Phy.	287961
2	Orthopedic	198489
3	Cardiovascular / Pulmonary	160867
4	General Medicine	120978
5	Neurology	110677
6	Gastroenterology	80347
7	Radiology	74969
8	Obstetrics / Gynecology	72589
9	Urology	63419
10	SOAP / Chart / Progress Notes	59558
11	Neurosurgery	54233
12	Discharge Summary	43103
13	Psychiatry / Psychology	42972
14	ENT - Otolaryngology	42032

```
total_word_count = df['transcription'].str.split().str.len().sum()
print(f'The word count of all transcription is: {int(total_word_count)}')
```

```
The word count of all transcription is: 139368
```

18	Pediatrics - Neonatal	30724

1.3 | Sample Size of Each Medical Specialty

```
vc(df, 'medical_specialty')
```

```
STATUS: Value counts of "medical_specialty"...
```

medical_specialty	count	percentage (%)
Discharge Summary	108	36.000000
ENT - Otolaryngology	98	32.666667
Neurosurgery	94	31.333333

1.4 | General Cleaning

```
# to print data shape
print(f'data shape is: {df.shape}')

# to identify the null values by descending order
df.isnull().sum().sort_values(ascending = False)
```

```
data shape is: (300, 5)
keywords             56
transcription         2
description           0
medical_specialty     0
sample_name           0
dtype: int64
```

One important detail is that I found out there are 2 rows containing no transcription. They should be removed as transcription is our only predictors in this text classification task.

```
# to remove transcription rows that is empty
df = df[df['transcription'].notna()]
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 298 entries, 2656 to 3994
Data columns (total 5 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   description        298 non-null    object
 1   medical_specialty  298 non-null    object
 2   sample_name        298 non-null    object
 3   transcription      298 non-null    object
 4   keywords           242 non-null    object
dtypes: object(5)
memory usage: 14.0+ KB
```

After dropping the null values, there are no null values for the transcription attribute.

```
# drop redundant columns
df =df.drop(['description','sample_name','keywords'], axis=1)
df.head(1000)
```

	medical_specialty	transcription
		TITLE OF OPERATION: A complex closure and debridement of wound. INDICATION FOR

The target labels (or the topic) is the 'medical_specialty' attribute. Now, let's identify how is the value counts of the target labels, and as well visualise it in a bar chart. In order to visualise in matplotlib, function of flattening list is defined in order to put the target value counts into the matplotlib function.

month-old infant, born premature with intraventricular hemorrhage...

The target labels is quite balanced

endotracheal tube anesthesia. DEVICES: Pectoral ventricular catheter with a

2.0 | Text Normalisation

Data normalisation will be conducted for the trascription. One of the reasons is to convert the transcript into standard format, which important for data extraction later. In this data normalisation task, following task will be executed, which are:

1. Lowe Case
2. Removing punctuation and numbers
3. Tokenisation of the transcription
4. Lemmatisation
5. Remove Stop Words

REASON FOR TRANSFER:, Need for cardiac catheterization done at ABCD., TRANSFER

2.1 | Lower Case

3991	Discharge Summary	C4-5 right with left radiculopathy.,2. Moderate stenosis C5-6.,OPERATION: , On 06/25/07, anterior

```
# To convert transcription into lowercase
def lower(df, attribute):
    df.loc[:,attribute] = df[attribute].apply(lambda x : str.lower(x))
    return df
df = lower(df,'transcription')
df.head(1000)
```

	medical_specialty	transcription
2656	Neurosurgery	title of operation:, a complex closure and debridement of wound.,indication for surgery:, the patient is a 26-year-old female with a long history of shunt and hydrocephalus presenting with a draining wound in the right upper quadrant, just below the...
2657	Neurosurgery	title of operation: , placement of right new ventriculoperitoneal (vp) shunts strata valve and to removal of right frontal ommaya reservoir, indication for surgery:, the patient is a 2 month old infant...

2.2 | Remove Punctuation and Numbers

```
# To remove transcription punctuation and numbers

warnings.filterwarnings('ignore')
def remove_punc_num(df, attribute):
    df.loc[:,attribute] = df[attribute].apply(lambda x : " ".join(re.findall('[\w]+',x)))
    df[attribute] = df[attribute].str.replace('\d+', '')
    return df
df =remove_punc_num(df, 'transcription')
df_no_punc =df.copy()
df.head(1000)
```

	medical_specialty	transcription
2656	Neurosurgery	title of operation a complex closure and debridement of wound indication for surgery the patient is a year old female with a long history of shunt and hydrocephalus presenting with a draining wound in the right upper quadrant just below the costal ma...
2657	Neurosurgery	title of operation placement of right new ventriculoperitoneal vp shunts strata valve and to removal of right frontal ommaya reservoir indication for surgery the patient is a month old infant born premature with intraventricular hemorrhage and ommaya...
2658	Neurosurgery	preoperative diagnosis aqueductal stenosis postoperative diagnosis aqueductal stenosis title of procedure endoscopic third ventriculostomy anesthesia general endotracheal tube anesthesia devices bactiseal ventricular catheter with an aesculap burr hol...
2661	Neurosurgery	procedure placement of left ventriculostomy via twist drill preoperative diagnosis massive intraventricular hemorrhage with hydrocephalus and increased intracranial pressure postoperative diagnosis massive intraventricular hemorrhage with hydrocephalu...
2662	Neurosurgery	preoperative diagnoses increased intracranial pressure and cerebral edema due to severe brain injury postoperative diagnoses increased intracranial pressure and cerebral edema due to severe brain injury procedure burr hole and insertion of external ve...
...	...	...
3984	Discharge Summary	discharge diagnoses brca mutation history of present illness the patient is a year old with a brca mutation her sister died of breast cancer at age and her daughter had breast cancer at age physical examination the chest was clear the abdomen was...
3985	Discharge Summary	chief complaint decreased ability to perform daily living activities secondary to exacerbation of chronic back pain history of present illness the patient is a year old white male who was admitted with acute back pain the patient reports that he had ...
3986	Discharge Summary	reason for transfer need for cardiac catheterization done at abcd transfer diagnoses coronary artery disease chest pain history of diabetes history of hypertension history of obesity a cm lesion in the medial aspect of the right parietal lobe ...
3991	Discharge Summary	final diagnoses herniated nucleuses pulposus c greater than c left greater than c right with left radiculopathy moderate stenosis c operation on anterior cervical discectomy and fusions c c c using bengal cages and slimlock plate c to c in...

2.3 | Tokenisation

```
# to tokenise transcription

# import nltk
tk =WhitespaceTokenizer()
```

```
def tokenise(df, attribute):
    df['tokenised'] = df.apply(lambda row: tk.tokenize(str(row[attribute])), axis=1)
    return df
df =tokenise(df, 'transcription')
df_experiment =df.copy()
df.head(1000)
```

	medical_specialty	transcription	tokenised
2656	Neurosurgery	title of operation a complex closure and debridement of wound indication for surgery the patient is a year old female with a long history of shunt and hydrocephalus presenting with a draining wound in the right upper quadrant just below the costal ma...	[title, of, operation, a, complex, closure, and, debridement, of, wound, indication, for, surgery, the, patient, is, a, year, old, female, with, a, long, history, of, shunt, and, hydrocephalus, presenting, with, a, draining, wound, in, the, right, upp...
2657	Neurosurgery	title of operation placement of right new ventriculoperitoneal vp shunts strata valve and to removal of right frontal ommaya reservoir indication for surgery the patient is a month old infant born premature with intraventricular hemorrhage and ommaya...	[title, of, operation, placement, of, right, new, ventriculoperitoneal, vp, shunts, strata, valve, and, to, removal, of, right, frontal, ommaya, reservoir, indication, for, surgery, the, patient, is, a, month, old, infant, born, premature, with, intra...
2658	Neurosurgery	preoperative diagnosis aqueductal stenosis postoperative diagnosis aqueductal stenosis title of procedure endoscopic third ventriculostomy anesthesia general endotracheal tube anesthesia devices bactiseal ventricular catheter with an aesculap burr hol...	[preoperative, diagnosis, aqueductal, stenosis, postoperative, diagnosis, aqueductal, stenosis, title, of, procedure, endoscopic, third, ventriculostomy, anesthesia, general, endotracheal, tube, anesthesia, devices, bactiseal, ventricular, catheter, w...
2661	Neurosurgery	procedure placement of left ventriculostomy via twist drill preoperative diagnosis massive intraventricular hemorrhage with hydrocephalus and increased intracranial pressure postoperative diagnosis massive intraventricular hemorrhage with hydrocephalu...	[procedure, placement, of, left, ventriculostomy, via, twist, drill, preoperative, diagnosis, massive, intraventricular, hemorrhage, with, hydrocephalus, and, increased, intracranial, pressure, postoperative, diagnosis, massive, intraventricular, hemo...
2662	Neurosurgery	preoperative diagnoses increased intracranial pressure and cerebral edema due to severe brain injury postoperative diagnoses increased intracranial pressure and cerebral edema due to severe brain injury procedure burr hole and insertion of external ve...	[preoperative, diagnoses, increased, intracranial, pressure, and, cerebral, edema, due, to, severe, brain, injury, postoperative, diagnoses, increased, intracranial, pressure, and, cerebral, edema, due, to, severe, brain, injury, procedure, burr, hole...
...	...	...	...
3984	Discharge Summary	discharge diagnoses brca mutation history of present illness the patient is a year old with a brca mutation her sister died of breast cancer at age and her daughter had breast cancer at age physical examination the chest was clear the abdomen was...	[discharge, diagnoses, brca, mutation, history, of, present, illness, the, patient, is, a, year, old, with, a, brca, mutation, her, sister, died, of, breast, cancer, at, age, and, her, daughter, had, breast, cancer, at, age, physical, examination, the...
3985	Discharge Summary	chief complaint decreased ability to perform daily living activities secondary to exacerbation of chronic back pain history of present illness the patient is a year old white male who was admitted with acute back pain the patient reports that he had ...	[chief, complaint, decreased, ability, to, perform, daily, living, activities, secondary, to, exacerbation, of, chronic, back, pain, history, of, present, illness, the, patient, is, a, year, old, white, male, who, was, admitted, with, acute, back, pai...
3986	Discharge Summary	reason for transfer need for cardiac catheterization done at abcd transfer diagnoses coronary artery disease chest pain history of diabetes history of hypertension history of obesity a cm lesion in the medial aspect of the right parietal lobe ...	[reason, for, transfer, need, for, cardiac, catheterization, done, at, abcd, transfer, diagnoses, coronary, artery, disease, chest, pain, history, of, diabetes, history, of, hypertension, history, of, obesity, a, cm, lesion, in, the, medial, aspect, o...
3991	Discharge Summary	final diagnoses herniated nucleuses pulposus c greater than c left greater than c right with left radiculopathy moderate stenosis c operation on anterior cervical discectomy and fusions c c c using bengal cages and slimlock plate c to c in...	[final, diagnoses, herniated, nucleuses, pulposus, c, greater, than, c, left, greater, than, c, right, with, left, radiculopathy, moderate, stenosis, c, operation, on, anterior, cervical, discectomy, and, fusions, c, c, c, using, bengal, cages, and, s...
		admitting diagnosis abscess with cellulitis left foot	[admitting, diagnosis, abscess, with, cellulitis

2.4 | Stemming

```
from nltk.stem.snowball import SnowballStemmer
def stemming(df, attribute):
    # Use English stemmer.
    stemmer = SnowballStemmer("english")
    df['stemmed'] = df[attribute].apply(lambda x: [stemmer.stem(y) for y in x]) # Stem every word.
    return df
df =stemming(df_experiment, 'tokenised')
df.head(1000)
```

	medical_specialty	transcription	tokenised	stemmed
2656	Neurosurgery	title of operation a complex closure and debridement of wound indication for surgery the patient is a year old female with a long history of shunt and hydrocephalus presenting with a draining wound in the right upper quadrant just below the costal ma...	[title, of, operation, a, complex, closure, and, debridement, of, wound, indication, for, surgery, the, patient, is, a, year, old, female, with, a, long, history, of, shunt, and, hydrocephalus, presenting, with, a, draining, wound, in, the, right, upp...	[titl, of, oper, a, complex, closur, and, debrid, of, wound, indic, for, surgeri, the, patient, is, a, year, old, femal, with, a, long, histori, of, shunt, and, hydrocephalus, present, with, a, drain, wound, in, the, right, upper, quadrant, just, belo...
2657	Neurosurgery	title of operation placement of right new ventriculoperitoneal vp shunts strata valve and to removal of right frontal ommaya reservoir indication for surgery the patient is a month old infant born premature with intraventricular hemorrhage and ommaya...	[title, of, operation, placement, of, right, new, ventriculoperitoneal, vp, shunts, strata, valve, and, to, removal, of, right, frontal, ommaya, reservoir, indication, for, surgery, the, patient, is, a, month, old, infant, born, premature, with, intra...	[titl, of, oper, placement, of, right, new, ventriculoperiton, vp, shunt, strata, valv, and, to, remov, of, right, frontal, ommaya, reservoir, indic, for, surgeri, the, patient, is, a, month, old, infant, born, prematur, with, intraventricular, hemorr...
2658	Neurosurgery	preoperative diagnosis aqueductal stenosis postoperative diagnosis aqueductal stenosis title of procedure endoscopic third ventriculostomy anesthesia	[preoperative, diagnosis, aqueductal, stenosis, postoperative, diagnosis, aqueductal, stenosis, title, of, procedure, endoscopic, third, ventriculostomy, anesthesia	[preoper, diagnosi, aqueduct, stenosi, postop, diagnosi, aqueduct, stenosi, titl, of, procedur, endoscop, third, ventriculostomi, anesthesia, general, endotrach, tube

2.5 | Stop Words Removal

Removing stop words from the feature space, otherwise it will affect the classifier performance as the collection frequency is often high

	preoperative diagnosis massive	preoperative, diagnosis,	preoper, diagnosi, massiv,

```
# Showing the list of the English stop words, it has a number of 179 stop words in this list
import nltk
nltk.download('stopwords')
stop = stopwords.words('english')
print(f"There are {len(stop)} stop words \n")
print(stop)
```

```
There are 179 stop words

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd",
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

	procedure burr hole and	cerebral, edema, due, to,	injur, procedur, burr, hole, and, insert, of, ventri

```
# Removing stop words
def remove_stop_words(df, attribute):
    stop = stopwords.words('english')
    df['stemmed_without_stop'] = df[attribute].apply(lambda x: ' '.join([word for word in x if word not in (stop)
    return df
df = remove_stop_words(df, 'stemmed')
df.head(1000)
```

	medical_specialty	transcription	tokenised	stemmed	stemmed_without_stop
2656	Neurosurgery	title of operation a complex closure and debridement of wound indication for surgery the patient is a year old female with a long history of shunt and hydrocephalus presenting with a draining wound in the right upper quadrant just below the costal ma...	[title, of, operation, a, complex, closure, and, debridement, of, wound, indication, for, surgery, the, patient, is, a, year, old, female, with, a, long, history, of, shunt, and, hydrocephalus, presenting, with, a, draining, wound, in, the, right, upp...	[titl, of, oper, a, complex, closur, and, debrid, of, wound, indic, for, surgeri, the, patient, is, a, year, old, femal, with, a, long, histori, of, shunt, and, hydrocephalus, present, with, a, drain, wound, in, the, right, upper, quadrant, just, belo...	titl oper complex closur debrid wound indic surgeri patient year old femal long histori shunt hydrocephalus present drain wound right upper quadrant costal margin lanc general surgeri resolv howev continu drain evid fever crp normal shunt ct normal th...
2657	Neurosurgery	title of operation placement of right new ventriculoperitoneal vp shunts strata valve and to removal of right frontal ommaya reservoir indication for surgery the patient is a month old infant born premature with intraventricular hemorrhage and ommaya...	[title, of, operation, placement, of, right, new, ventriculoperitoneal, vp, shunts, strata, valve, and, to, removal, of, right, frontal, ommaya, reservoir, indication, for, surgery, the, patient, is, a, month, old, infant, born, premature, with, intra...	[titl, of, oper, placement, of, right, new, ventriculoperiton, vp, shunt, strata, valv, and, to, remov, of, right, frontal, ommaya, reservoir, indic, for, surgeri, the, patient, is, a, month, old, infant, born, prematur, with, intraventricular, hemorr...	titl oper placement right new ventriculoperiton vp shunt strata valv remov right frontal ommaya reservoir indic surgeri patient month old infant born prematur intraventricular hemorrhag ommaya reservoir recommend remov replac new vp shunt preop diagno...
2658	Neurosurgery	preoperative diagnosis aqueductal stenosis postoperative diagnosis aqueductal stenosis title of procedure endoscopic third ventriculostomy anesthesia general endotracheal tube anesthesia devices bactiseal ventricular catheter with an aesculap burr hol...	[preoperative, diagnosis, aqueductal, stenosis, postoperative, diagnosis, aqueductal, stenosis, title, of, procedure, endoscopic, third, ventriculostomy, anesthesia, general, endotracheal, tube, anesthesia, devices, bactiseal, ventricular, catheter, w...	[preoper, diagnosi, aqueduct, stenosi, postop, diagnosi, aqueduct, stenosi, titl, of, procedur, endoscop, third, ventriculostomi, anesthesia, general, endotrach, tube, anesthesia, devic, bactis, ventricular, cathet, with, an, aesculap, burr, hole, por...	preoper diagnosi aqueduct stenosi postop diagnosi aqueduct stenosi titl procedur endoscop third ventriculostomi anesthesia general endotrach tube anesthesia devic bactis ventricular cathet aesculap burr hole port skin prepar chloraprep complic none sp...
2661	Neurosurgery	procedure placement of left ventriculostomy via twist drill preoperative diagnosis massive intraventricular hemorrhage with hydrocephalus and increased intracranial pressure postoperative diagnosis massive intraventricular hemorrhage with hydrocephalu...	[procedure, placement, of, left, ventriculostomy, via, twist, drill, preoperative, diagnosis, massive, intraventricular, hemorrhage, with, hydrocephalus, and, increased, intracranial, pressure, postoperative, diagnosis, massive, intraventricular, hemo...	[procedur, placement, of, left, ventriculostomi, via, twist, drill, preoper, diagnosi, massiv, intraventricular, hemorrhag, with, hydrocephalus, and, increas, intracrani, pressur, postop, diagnosi, massiv, intraventricular, hemorrhag, with, hydrocepha...	procedur placement left ventriculostomi via twist drill preoper diagnosi massiv intraventricular hemorrhag hydrocephalus increas intracrani pressur postop diagnosi massiv intraventricular hemorrhag hydrocephalus increas intracrani pressur indic proced...
2662	Neurosurgery	preoperative diagnoses increased intracranial pressure and cerebral edema due to severe brain injury postoperative diagnoses increased intracranial pressure and cerebral edema due to severe brain injury procedure burr hole and insertion of external ve...	[preoperative, diagnoses, increased, intracranial, pressure, and, cerebral, edema, due, to, severe, brain, injury, postoperative, diagnoses, increased, intracranial, pressure, and, cerebral, edema, due, to, severe, brain, injury, procedure, burr, hole...	[preoper, diagnos, increas, intracrani, pressur, and, cerebr, edema, due, to, sever, brain, injuri, postop, diagnos, increas, intracrani, pressur, and, cerebr, edema, due, to, sever, brain, injuri, procedur, burr, hole, and, insert, of, extern, ventri...	preoper diagnos increas intracrani pressur cerebr edema due sever brain injuri postop diagnos increas intracrani pressur cerebr edema due sever brain injuri procedur burr hole insert extern ventricular drain cathet anesthesia bedsid sedat procedur sca...

...	...	...	...	...	...
3984	Discharge Summary	discharge diagnoses brca mutation history of present illness the patient is a year old with a brca mutation her sister died of breast cancer at age and her daughter had breast cancer at age physical examination the chest was clear the abdomen was...	[discharge, diagnoses, brca, mutation, history, of, present, illness, the, patient, is, a, year, old, with, a, brca, mutation, her, sister, died, of, breast, cancer, at, age, and, her, daughter, had, breast, cancer, at, age, physical, examination, the...	[discharg, diagnos, brca, mutat, histori, of, present, ill, the, patient, is, a, year, old, with, a, brca, mutat, her, sister, die, of, breast, cancer, at, age, and, her, daughter, had, breast, cancer, at, age, physic, examin, the, chest, was, clear, ...	discharg diagnos brca mutat histori present ill patient year old brca mutat sister die breast cancer age daughter breast cancer age physic examin chest clear abdomen nontend pelvic examin show mass heart murmur hospit cours patient underw surgeri day ...
3985	Discharge Summary	chief complaint decreased ability to perform daily living activities secondary to exacerbation of chronic back pain history of present illness the patient is a year old white male who was admitted with acute back pain the patient reports that he had ...	[chief, complaint, decreased, ability, to, perform, daily, living, activities, secondary, to, exacerbation, of, chronic, back, pain, history, of, present, illness, the, patient, is, a, year, old, white, male, who, was, admitted, with, acute, back, pai...	[chief, complaint, decreas, abil, to, perform, daili, live, activ, secondari, to, exacerb, of, chronic, back, pain, histori, of, present, ill, the, patient, is, a, year, old, white, male, who, was, admit, with, acut, back, pain, the, patient, report, ...	chief complaint decreas abil perform daili live activ secondari exacerb chronic back pain histori present ill patient year old white male admit acut back pain patient report chronic problem back pain approxim year gotten progress wors last year patien...
		reason for transfer need for cardiac catheterization done at	[reason, for, transfer, need, for, cardiac, catheterization, done, at,	[reason, for, transfer, need, for, cardiac, catheter, done, at,	reason transfer need cardiac catheter done abcd transfer diagnos

After the 5 data normalisation steps, each transcription record is now in a standard format, which is ready for the n-gram features extraction later. Hence, we should use the attribute 'stemmed_withou_stop' as the predictor attribute and drop other redundant attributes, namely 'transcription', 'tokenized_transcription' and 'stemmed'.

		history of obesity a cm	hypertension, history,	obes, a, cm, lesion,	decondit consult

```
df =df.drop(['transcription','stemmed', 'tokenised'], axis=1)
df.head()
```

	medical_specialty	stemmed_without_stop
2656	Neurosurgery	titl oper complex closur debrid wound indic surgeri patient year old femal long histori shunt hydrocephalus present drain wound right upper quadrant costal margin lanc general surgeri resolv howev continu drain evid fever crp normal shunt ct normal th...
2657	Neurosurgery	titl oper placement right new ventriculoperiton vp shunt strata valv remov right frontal ommaya reservoir indic surgeri patient month old infant born prematur intraventricular hemorrhag ommaya reservoir recommend remov replac new vp shunt preop diagno...
2658	Neurosurgery	preoper diagnosi aqueduct stenosi postop diagnosi aqueduct stenosi titl procedur endoscop third ventriculostomi anesthesia general endotrach tube anesthesia devic bactis ventricular cathet aesculap burr hole port skin prepar chloraprep complic none sp...
2661	Neurosurgery	procedur placement left ventriculostomi via twist drill preoper diagnosi massiv intraventricular hemorrhag hydrocephalus increas intracrani pressur postop diagnosi massiv intraventricular hemorrhag hydrocephalus increas intracrani pressur indic proced...
2662	Neurosurgery	preoper diagnos increas intracrani pressur cerebr edema due sever brain injuri postop diagnos increas intracrani pressur cerebr edema due sever brain injuri procedur burr hole insert extern ventricular drain cathet anesthesia bedsid sedat procedur sca...

```
total_word_count_normalised = df['stemmed_without_stop'].str.split().str.len().sum()
print(f'The word count of transcription after normalised is: {int(total_word_count_normalised)}')
print(f'{round((total_word_count - total_word_count_normalised)/total_word_count*100, 2)}% less word')
```

```
The word count of transcription after normalised is: 83160
40.33% less word
```

```
le = preprocessing.LabelEncoder()
le.fit(df['medical_specialty'])
df['encoded_target'] = le.transform(df['medical_specialty'])
```

```
df.head()
```

	medical_specialty	stemmed_without_stop	encoded_target
2656	Neurosurgery	titl oper complex closur debrid wound indic surgeri patient year old femal long histori shunt hydrocephalus present drain wound right upper quadrant costal margin lanc general surgeri resolv howev continu drain evid fever crp normal shunt ct normal th...	2
2657	Neurosurgery	titl oper placement right new ventriculoperiton vp shunt strata valv remov right frontal ommaya reservoir indic surgeri patient month old infant born prematur intraventricular hemorrhag ommaya reservoir recommend remov replac new vp shunt preop diagno...	2
2658	Neurosurgery	preoper diagnosi aqueduct stenosi postop diagnosi aqueduct stenosi titl procedur endoscop third ventriculostomi anesthesia general endotrach tube anesthesia devic bactis ventricular cathet aesculap burr hole port skin prepar chloraprep complic none sp...	2
2661	Neurosurgery	procedur placement left ventriculostomi via twist drill preoper diagnosi massiv intraventricular hemorrhag hydrocephalus increas intracrani pressur postop diagnosi massiv intraventricular hemorrhag hydrocephalus increas intracrani pressur indic proced...	2
2662	Neurosurgery	preoper diagnos increas intracrani pressur cerebr edema due sever brain injuri postop diagnos increas intracrani pressur cerebr edema due sever brain injuri procedur burr hole insert extern ventricular drain cathet anesthesia bedsid sedat	2

3.0 | Text N-Gram Feature Extraction

We will use sklearn class 'CountVectoriser' to extract different n-grams features. In order to do so, the transcription should be converted into a list format, rather than a dataframe. For the purpose of converting into a flat list (i.e., there is no inner list), the function of 'flat_list' that defined above is used.

```
# function to flatten one list
def flat_list(unflat_list):
    flatted = [item for sublist in unflat_list for item in sublist]
    return flatted

def to_list(df, attribute):
    # Select the normalised transcript column
    df_transcription = df[[attribute]]
    # To convert the attribute into list format, but it has inner list. So it cannot put into the CountVectoriser
    unflat_list_transcription = df_transcription.values.tolist()
    # Let's use back the function defined above, "flat_list", to flatten the list
    flat_list_transcription = flat_list(unflat_list_transcription)
    return flat_list_transcription
flat_list_transcription = to_list(df, 'stemmed_without_stop')
```

3.1 | Extract 5 Types of N-Gram

CountVectorizer is used to convert a collection of transcript documents to a matrix of n-gram features. To explain the ngram_range, all values of n such such that min_n <= n <= max_n will be used. For example an ngram_range of (1, 1) means only unigrams, (1, 2) means unigrams and bigrams, and (2, 2) means only bigrams.

```
n_gram_features ={'unigram':(1,1),'unigram_bigram':(1,2),'bigram':(2,2),\
       'bigram_trigram':(2,3),'trigram':(3,3)}
feature_name=[]
temp=[]
for key, values in n_gram_features.items():
    temp.append(key)
```

```
    feature_name.append(key)
temp
```

```
['unigram', 'unigram_bigram', 'bigram', 'bigram_trigram', 'trigram']
```

```
def generate_n_gram_features(flat_list_transcription):
    temp=[]
    for key, values in n_gram_features.items():
        vectorizer = CountVectorizer(ngram_range=values)
        vectorizer.fit(flat_list_transcription)
        temp.append(vectorizer.transform(flat_list_transcription))
    return temp
temp = generate_n_gram_features(flat_list_transcription)
```

3.2 | Dimension of Each Feature Vector

```
dataframes = {'unigram':temp[0],
              'unigram_bigram':temp[1],
              'bigram':temp[2],
              'bigram_trigram':temp[3],
              'trigram':temp[4]}
feature_vector = [] ; feature_vector_shape = []
for key in dataframes:
    feature_vector.append(key)
    feature_vector_shape.append(dataframes[key].shape)

n_gram_df = pd.DataFrame({'N-Gram Feature Vector':feature_vector, 'Data Dimension':feature_vector_shape})
n_gram_df
```

	N-Gram Feature Vector	Data Dimension
0	unigram	(298, 5604)
1	unigram_bigram	(298, 54038)
2	bigram	(298, 48434)
3	bigram_trigram	(298, 115329)
4	trigram	(298, 66895)

After the feature extraction process, 5 kinds of n-gram features are extracted. It is interesting to notice that when the number of 'n' getting higher (i.e, n=1:unigram, n=2:bigram, n=3:trigram), there is a higer number of columns. This is due to it is getting harder to find similar features that can be stored in similar column when it has a longer connected words as one featuer. If the feature is unique, it will automatically append additional column to store the feaure.

```
# to retrieve a unigram feature vector
dataframes['unigram']
```

```
<298x5604 sparse matrix of type '<class 'numpy.int64'>'
        with 48505 stored elements in Compressed Sparse Row format>
```

4.0 | Text Classification Modelling

```
from sklearn.metrics import accuracy_score,f1_score,precision_score,recall_score
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
import warnings
from sklearn.experimental import enable_halving_search_cv  # noqa
from sklearn.model_selection import HalvingGridSearchCV
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
```

```
from sklearn.feature_selection import SelectFromModel
from sklearn.decomposition import PCA
from sklearn.metrics import classification_report

warnings.filterwarnings('ignore')
random_state_number =8888
df_target =df[['encoded_target']].values.ravel()


metrics = {
    'f1':[f1_score, 'f1_macro'],
    'precision': [precision_score, 'precision_macro'],
    'recall': [recall_score, 'recall_macro']
}

# get evaluation result

def get_performance(param_grid, base_estimator, dataframes):
    df_name_list =[]; best_estimator_list=[]; best_score_list=[]; test_predict_result_list=[];
    metric_list = [];

    for df_name, df in dataframes.items():

        X_train, X_test, y_train, y_test = train_test_split(df, df_target, test_size=0.2, random_state=random_sta
        for _, metric_dict in metrics.items():
            sh = HalvingGridSearchCV(base_estimator, param_grid, cv=5, scoring=metric_dict[1],random_state=random
                                     factor=2).fit(X_train, y_train)

            best_estimator = sh.best_estimator_
            clf = best_estimator.fit(X_train, y_train)
            prediction = clf.predict(X_test)
            test_predict_result = metric_dict[0](y_test, prediction, average='macro')

            df_name_list.append(df_name) ; best_estimator_list.append(best_estimator) ;
            best_score_list.append(sh.best_score_) ;
            test_predict_result_list.append(test_predict_result) ;metric_list.append(metric_dict[1])


    model_result = pd.DataFrame({'Vector':df_name_list,'Metric':metric_list,
                                 'Calibrated Estimator':best_estimator_list,
                                 'Best CV Metric Score':best_score_list, 'Test Predict Metric Score': test_predict_
    return model_result
```

4.1 | Visualising Classification Prediction

```
font = {'family' : 'Tahoma',
        'weight' : 'bold',
        'size'   : 12}
matplotlib.rc('font', **font)

def vis_classification(vector_type = 'unigram', estimator = KNeighborsClassifier(n_neighbors=9)):
    pca = PCA(n_components=2)
    df1 = pca.fit_transform(dataframes[vector_type].todense())
    X_train, X_test, y_train, y_test = train_test_split(df1, df_target, test_size=0.2, random_state=random_state_

    # get training set
    df2 = pd.DataFrame({'pca1':X_train[:,1], 'pca2': X_train[:,0], 'y':le.inverse_transform(y_train)})
    min_1, max_1 = df2['pca1'].min(), df2['pca1'].max()
    min_2, max_2 = df2['pca2'].min(), df2['pca2'].max()

    # generate dimension reduced, but extended data
    pca1_range = np.linspace(min_1,max_1,30)
    pca2_range = np.linspace(min_2,max_2,30)
```

```
    # shuffle
    np.random.shuffle(pca1_range) ; np.random.shuffle(pca2_range)

    # to dataframe
    prediction_test = pd.DataFrame({'pca1':pca1_range, 'pca2':pca2_range})

    best_estimator = estimator

    # fit training set and predict extended data
    clf = best_estimator.fit(X_train, y_train)

    fig, axs = plt.subplots(nrows = 1, ncols = 2, figsize=(15,6))
    cmap = plt.cm.get_cmap('tab10', 4)
    fig.suptitle(f"Visualising {type(estimator).__name__} on {vector_type.capitalize()} Vector", fontsize=14,font


    def plot_scatter(ax, predictor_set, target, title):

        # plot area classifier
        clf = best_estimator.fit(X_train, y_train)
        axs[0].tricontourf(X_train[:,0], X_train[:,1], clf.predict(X_train), levels=np.arange(-0.5, 4), zorder=10

        axs[1].tricontourf(X_test[:,0], X_test[:,1], clf.predict(X_test), levels=np.arange(-0.5, 4), zorder=10, a

        # plot scatter
        df3 = pd.DataFrame({'pca1':predictor_set[:,1], 'pca2': predictor_set[:,0], 'y':le.inverse_transform(targe
        for y_label in df3['y'].unique():
            df_filter = df3[df3['y']==y_label]
            ax.scatter(df_filter['pca1'], df_filter['pca2'], alpha=1,label=f"{y_label}")
        ax.legend()
        ax.set_title(f'{title} ({predictor_set.shape[0]} Samples)',fontweight='bold')
    plot_scatter(axs[0], X_train, y_train, 'Training Set')
    plot_scatter(axs[1], X_test, y_test, 'Testing Set')
    axs[0].sharey(axs[1])
    return plt.show()

param_grid = {'max_depth': [None,30,32,35,37,38,39,40],'min_samples_split': [2,150,170,180,190,200]}
base_estimator = RandomForestClassifier(random_state=random_state_number)
rfc_result = get_performance(param_grid, base_estimator, dataframes)
rfc_result
```

	Vector	Metric	Calibrated Estimator	Best CV Metric Score	Test Predict Metric Score
0	unigram	f1_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507),\n DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502),\n DecisionTreeClassifier(max_depth=3...	0.815681	0.902071
1	unigram	precision_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507),\n DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502),\n DecisionTreeClassifier(max_depth=3...	0.860206	0.909018
2	unigram	recall_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507),\n DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502),\n DecisionTreeClassifier(max_depth=3...	0.826720	0.912037
3	unigram_bigram	f1_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507),\n DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502),\n DecisionTreeClassifier(max_depth=3...	0.766749	0.868443
4	unigram_bigram	precision_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507),\n DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502),\n DecisionTreeClassifier(max_depth=3...	0.812359	0.881579
5	unigram_bigram	recall_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507),\n DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502),\n DecisionTreeClassifier(max_depth=3...	0.781481	0.884259
6	bigram	f1_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507),\n DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502),\n DecisionTreeClassifier(max_depth=3...	0.674685	0.851420
7	bigram	precision_macro	(DecisionTreeClassifier(max_depth=35, max_features='auto',\n random_state=1985925507),\n DecisionTreeClassifier(max_depth=35, max_features='auto',\n random_state=1459224502),	0.806748	0.855724

```
def get_best_vector_clf(knn_result):

    temp = knn_result[knn_result['Metric'] =='f1_macro']
    temp2 = temp.iloc[temp['Best CV Metric Score'].idxmax()].to_frame().T
    best_vector = temp2['Vector'].values[0]
    best_clf = temp2['Calibrated Estimator'].values[0]\

    return best_vector, best_clf

best_vector, best_clf =  get_best_vector_clf(rfc_result)
#vis_classification(vector_type = best_vector, estimator = best_clf)
```

4.2 | Dimensionality Reduction

```
df_temp = rfc_result[rfc_result['Metric'] =='f1_macro']
# df_temp['Calibrated Estimator']
vector_rfc = df_temp[['Vector','Calibrated Estimator']].set_index('Vector').to_dict()['Calibrated Estimator']
vector_rfc
```

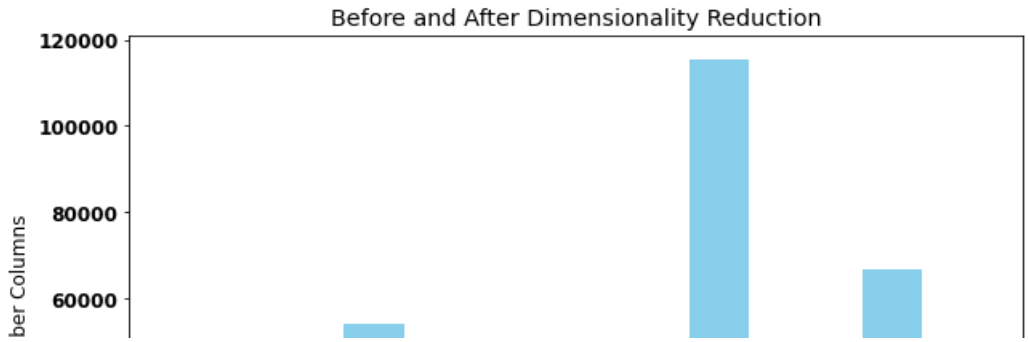
```
     {'unigram': RandomForestClassifier(max_depth=32, random_state=8888),
      'unigram_bigram': RandomForestClassifier(max_depth=32, random_state=8888),
      'bigram': RandomForestClassifier(max_depth=32, random_state=8888),
      'bigram_trigram': RandomForestClassifier(max_depth=30, random_state=8888),
      'trigram': RandomForestClassifier(max_depth=30, random_state=8888)}


supported_columns_dict = {}
for df_name, df in dataframes.items():
    X_train, X_test, y_train, y_test = train_test_split(dataframes[df_name], df_target, test_size=0.2, random_sta

    selector = SelectFromModel(estimator=vector_rfc[df_name]).fit(X_train, y_train)

    filter_columns = selector.get_support()
    dataframes[df_name] = dataframes[df_name][:, filter_columns]

shape_dim = [] ; df_names = []
for df_name, df in dataframes.items():
    shape_dim.append(df.shape)
    df_names.append(df_name)
n_gram_df_dim = pd.DataFrame({'N-Gram Feature Vector':df_names, 'Data Dimension':shape_dim})
n_gram_df_dim
```

	N-Gram Feature Vector	Data Dimension
0	unigram	(298, 974)
1	unigram_bigram	(298, 3074)
2	bigram	(298, 3334)
3	bigram_trigram	(298, 3551)
4	trigram	(298, 3139)

```
labels = n_gram_df_dim['N-Gram Feature Vector'].values
b4 = [shape[1] for shape in n_gram_df['Data Dimension'].values]
af = [shape[1] for shape in n_gram_df_dim['Data Dimension'].values]

x = np.arange(len(labels))  # the label locations
width = 0.35  # the width of the bars

fig, ax = plt.subplots(figsize=(10, 6))
rects1 = ax.bar(x - width/2, b4, width, label='Before Dimensionality Reduction', color='skyblue')
rects2 = ax.bar(x + width/2, af, width, label='After Dimensionality Reduction', color='lime')

# Add some text for labels, title and custom x-axis tick labels, etc.
ax.set_ylabel('Number Columns')
ax.set_title('Before and After Dimensionality Reduction')
ax.set_xticks(x, labels)
ax.set_xticklabels(ax.get_xticklabels(),rotation=30)
ax.legend()

ax.bar_label(rects1, padding=3)
ax.bar_label(rects2, padding=3)

fig.tight_layout()

plt.show()
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-48-14e5cc25f64d> in <module>
     13 ax.set_ylabel('Number Columns')
     14 ax.set_title('Before and After Dimensionality Reduction')
---> 15 ax.set_xticks(x, labels)
     16 ax.set_xticklabels(ax.get_xticklabels(),rotation=30)
     17 ax.legend()

3 frames
/usr/local/lib/python3.7/dist-packages/matplotlib/axis.py in set_ticks(self, ticks, minor)
   1766             else:
   1767                 self.set_view_interval(max(ticks), min(ticks))
-> 1768         if minor:
   1769             self.set_minor_locator(mticker.FixedLocator(ticks))
   1770             return self.get_minor_ticks(len(ticks))

ValueError: The truth value of an array with more than one element is ambiguous. Use a.any() or a.all()
```

SEARCH STACK OVERFLOW



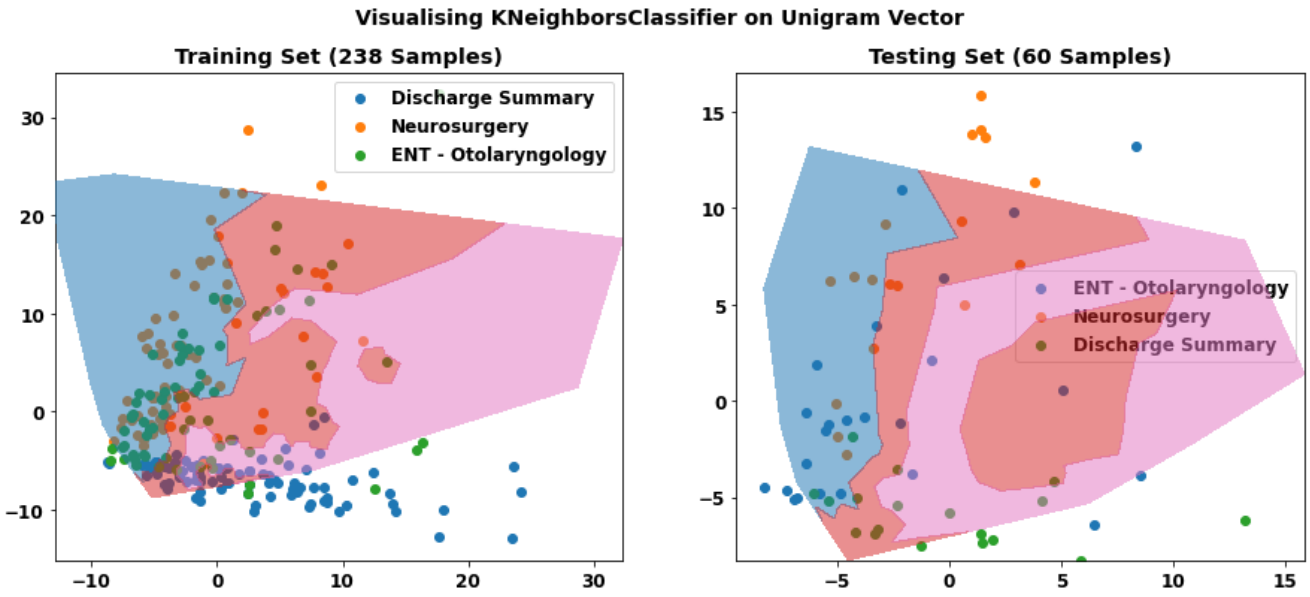

```
param_grid = {'n_neighbors': [5,7,9,11,13,15,17,19,21]}
base_estimator = KNeighborsClassifier()
knn_result = get_performance(param_grid, base_estimator, dataframes)
knn_result
```

	Vector	Metric	Calibrated Estimator	Best CV Metric Score	Test Predict Metric Score
0	unigram	f1_macro	KNeighborsClassifier(n_neighbors=9)	0.656608	0.821654
1	unigram	precision_macro	KNeighborsClassifier(n_neighbors=9)	0.805632	0.860119
2	unigram	recall_macro	KNeighborsClassifier(n_neighbors=9)	0.668915	0.833333
3	unigram_bigram	f1_macro	KNeighborsClassifier(n_neighbors=7)	0.589224	0.790888
4	unigram_bigram	precision_macro	KNeighborsClassifier(n_neighbors=9)	0.781045	0.880952
5	unigram_bigram	recall_macro	KNeighborsClassifier(n_neighbors=7)	0.612434	0.800926
6	bigram	f1_macro	KNeighborsClassifier(n_neighbors=17)	0.182196	0.153846
7	bigram	precision_macro	KNeighborsClassifier(n_neighbors=17)	0.180259	0.100000
8	bigram	recall_macro	KNeighborsClassifier(n_neighbors=17)	0.340741	0.333333
9	bigram_trigram	f1_macro	KNeighborsClassifier(n_neighbors=17)	0.182196	0.153846
10	bigram_trigram	precision_macro	KNeighborsClassifier(n_neighbors=17)	0.180259	0.100000
11	bigram_trigram	recall_macro	KNeighborsClassifier(n_neighbors=17)	0.340741	0.333333
12	trigram	f1_macro	KNeighborsClassifier(n_neighbors=17)	0.167877	0.153846
13	trigram	precision_macro	KNeighborsClassifier(n_neighbors=17)	0.112802	0.100000
14	trigram	recall_macro	KNeighborsClassifier(n_neighbors=17)	0.333333	0.333333

```
best_vector, best_clf =  get_best_vector_clf(knn_result)
vis_classification(vector_type = best_vector, estimator = best_clf)
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-39-e9d3d8f15280> in <module>
      1 best_vector, best_clf =  get_best_vector_clf(knn_result)
----> 2 vis_classification(vector_type = best_vector, estimator = best_clf)

<ipython-input-30-e552282b987c> in vis_classification(vector_type, estimator)
     51     plot_scatter(axs[0], X_train, y_train, 'Training Set')
     52     plot_scatter(axs[1], X_test, y_test, 'Testing Set')
---> 53     axs[0].sharey(axs[1])
     54     return plt.show()

AttributeError: 'AxesSubplot' object has no attribute 'sharey'
```

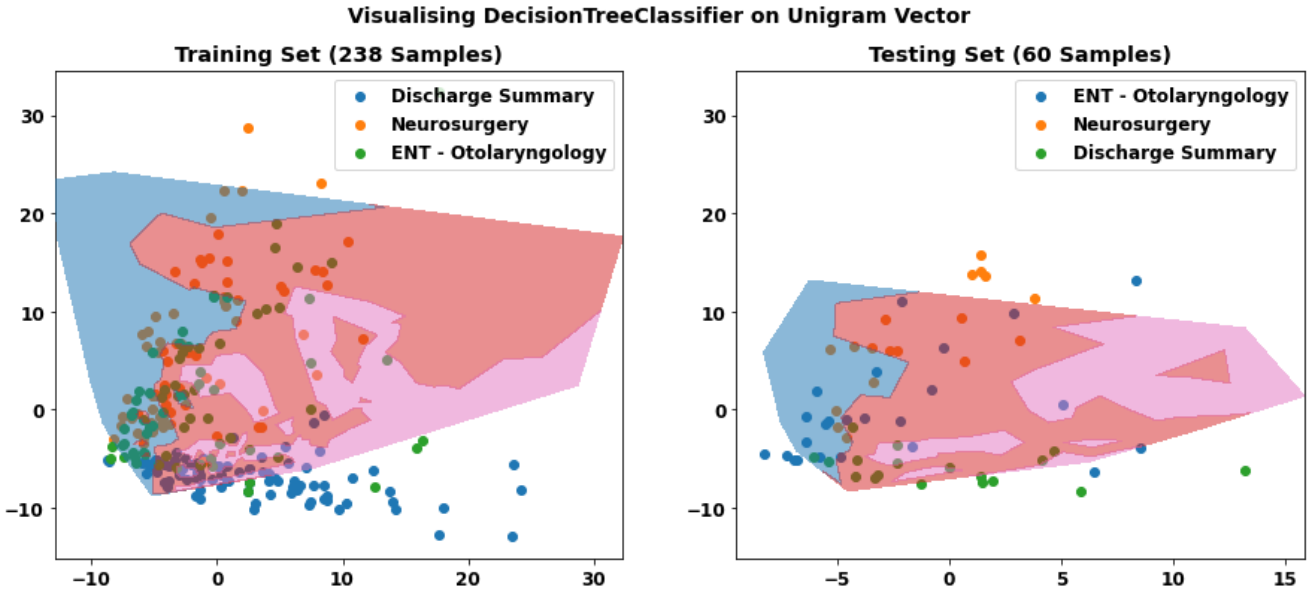
SEARCH STACK OVERFLOW




```
param_grid = {'max_depth': [None,4,6,7,8,30,32,35],'min_samples_split': [2,3,4,5,35,10,16,20]}
base_estimator = DecisionTreeClassifier(random_state=random_state_number)
dtc_result = get_performance(param_grid, base_estimator, dataframes)
dtc_result
```

	Vector	Metric	Calibrated Estimator	Best CV Metric Score	Test Predict Metric Score
0	unigram	f1_macro	DecisionTreeClassifier(max_depth=35, min_samples_split=5, random_state=8888)	0.846548	0.871630
1	unigram	precision_macro	DecisionTreeClassifier(min_samples_split=4, random_state=8888)	0.860370	0.881297
2	unigram	recall_macro	DecisionTreeClassifier(max_depth=8, min_samples_split=4, random_state=8888)	0.845132	0.875000
3	unigram_bigram	f1_macro	DecisionTreeClassifier(max_depth=6,	0.841640	0.870052

```
best_vector, best_clf =  get_best_vector_clf(dtc_result)
vis_classification(vector_type = best_vector, estimator = best_clf)
```



13	trigram	precision_macro	DecisionTreeClassifier(max_depth=8,	0.714302	0.602083

```
df_result = pd.concat([knn_result,
                       dtc_result,
                       rfc_result
                       ]
                      ).reset_index(drop=True)

df_result.groupby(['Metric']).max()
```

Metric	Vector	Best CV Metric Score	Test Predict Metric Score
f1_macro	unigram_bigram	0.846548	0.902071
precision_macro	unigram_bigram	0.860370	0.909018
recall_macro	unigram_bigram	0.845132	0.912037

4.3 | Obtain Best Classifier and Feature Vector

```
def get_best_result(df_result, metric_score):
    df_result_t = df_result[df_result.Metric== 'precision_macro']
    precision_macro_df = df_result_t.loc[df_result_t[metric_score].idxmax()].to_frame().T

    df_result_t = df_result[df_result.Metric== 'recall_macro']
    recall_macro_df = df_result_t.loc[df_result_t[metric_score].idxmax()].to_frame().T
```

```
    df_result_t = df_result[df_result.Metric== 'f1_macro']
    f1_macro_df = df_result_t.loc[df_result_t[metric_score].idxmax()].to_frame().T

    return pd.concat([precision_macro_df,recall_macro_df,f1_macro_df])


best_cv_result = get_best_result(df_result, 'Best CV Metric Score')
display(best_cv_result)
temp = best_cv_result[best_cv_result['Metric'] == 'f1_macro']
best_clf = temp['Calibrated Estimator'].values[0]
best_vector = temp['Vector'].values[0]
```

	Vector	Metric	Calibrated Estimator	Best CV Metric Score	Test Predict Metric Score
16	unigram	precision_macro	DecisionTreeClassifier(min_samples_split=4, random_state=8888)	0.86037	0.881297
17	unigram	recall_macro	DecisionTreeClassifier(max_depth=8, min_samples_split=4, random_state=8888)	0.845132	0.875
15	unigram	f1_macro	DecisionTreeClassifier(max_depth=35, min_samples_split=5, random_state=8888)	0.846548	0.87163

```
get_best_result(df_result, 'Test Predict Metric Score')
```

	Vector	Metric	Calibrated Estimator	Best CV Metric Score	Test Predict Metric Score
31	unigram	precision_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507), DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502), DecisionTreeClassifier(max_depth=3...	0.860206	0.909018
32	unigram	recall_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507), DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502), DecisionTreeClassifier(max_depth=3...	0.82672	0.912037
30	unigram	f1_macro	(DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1985925507), DecisionTreeClassifier(max_depth=32, max_features='auto',\n random_state=1459224502), DecisionTreeClassifier(max_depth=3...	0.815681	0.902071

4.4 | Evaluate on Each Class Labels

```
X_train, X_test, y_train, y_test = train_test_split(dataframes[best_vector], df_target, test_size=0.2, \
                                                    random_state=random_state_number)
clf = best_clf.fit(X_train, y_train)
y_test_pred= clf.predict(X_test)
target_names = ['Discharge Summary', 'ENT', 'Neurosurgery']
print(classification_report(y_test,y_test_pred,target_names=target_names))
```

```
                    precision    recall  f1-score   support

     Discharge Summary       1.00      0.89      0.94        18
                   ENT       0.90      0.79      0.84        24
          Neurosurgery       0.74      0.94      0.83        18

              accuracy                           0.87        60
             macro avg       0.88      0.88      0.87        60
          weighted avg       0.88      0.87      0.87        60
```

```
sample_predict = pd.DataFrame({'Actual Y Test': le.inverse_transform(y_test),'Best Prediction':le.inverse_transfo
sample_predict.head(20)
```

	Actual Y Test	Best Prediction
0	ENT - Otolaryngology	ENT - Otolaryngology
1	ENT - Otolaryngology	ENT - Otolaryngology
2	ENT - Otolaryngology	ENT - Otolaryngology
3	ENT - Otolaryngology	Neurosurgery
4	Neurosurgery	Neurosurgery
5	Neurosurgery	Neurosurgery
6	ENT - Otolaryngology	ENT - Otolaryngology
7	Discharge Summary	Discharge Summary
8	Neurosurgery	Neurosurgery
9	ENT - Otolaryngology	ENT - Otolaryngology
10	Discharge Summary	Discharge Summary
11	Neurosurgery	Neurosurgery
12	Neurosurgery	Neurosurgery
13	ENT - Otolaryngology	ENT - Otolaryngology
14	ENT - Otolaryngology	ENT - Otolaryngology
15	Discharge Summary	Discharge Summary
16	ENT - Otolaryngology	ENT - Otolaryngology
17	Neurosurgery	Neurosurgery
18	Discharge Summary	Discharge Summary
19	ENT - Otolaryngology	ENT - Otolaryngology