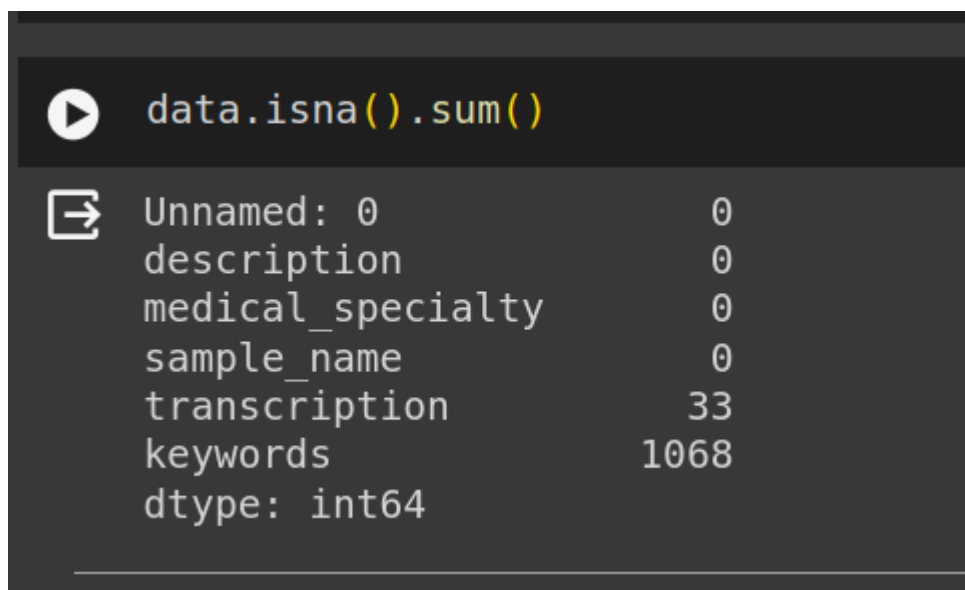


Assignment : Multi Label Classification

Data Understanding and Exploration:

The given data contains 5000 rows and 6 columns. Columns are namely description, medical_specialty, sample_name, transcription, keywords. Data contains some null values. The following image shows the column wise null values present in the dataset.



```
data.isna().sum()
Unnamed: 0      0
description      0
medical_specialty  0
sample_name      0
transcription    33
keywords        1068
dtype: int64
```

There are 40 unique values for the column medical_specialty. The distribution of the data between these classes is as follows.

Surgery	1088
Consult - History and Phy.	516
Cardiovascular / Pulmonary	371
Orthopedic	355
Radiology	273
General Medicine	259
Gastroenterology	224
Neurology	223
SOAP / Chart / Progress Notes	166
Urology	156
Obstetrics / Gynecology	155
Discharge Summary	108
ENT - Otolaryngology	96
Neurosurgery	94
Hematology - Oncology	90

Ophthalmology	83
Nephrology	81
Emergency Room Reports	75
Pediatrics - Neonatal	70
Pain Management	61
Psychiatry / Psychology	53
Office Notes	50
Podiatry	47
Dermatology	29
Cosmetic / Plastic Surgery	27
Dentistry	27
Letters	23
Physical Medicine - Rehab	21
Sleep Medicine	20
Endocrinology	19
Bariatrics	18
IME-QME-Work Comp etc.	16
Chiropractic	14
Rheumatology	10
Diets and Nutritions	10
Speech - Language	9
Autopsy	8
Lab Medicine - Pathology	8
Allergy / Immunology	7
Hospice - Palliative Care	6

The following diagram will help understand the distribution of the data in the provided classes.

Approach:

To classify the data based on the transcripts we can pass the transcripts to the BERT model. The BERT model will provide us with the context vector for each transcript. These obtained context vectors from the BERT model can then be passed to the fully connected layer. Fully connected layers will be responsible to correctly classify the data. Since we are performing multi-class classification hence cross entropy loss can be used to calculate the loss.

We can also use the LSTM to obtain the meaningful values from the given transcription.

