**NBME - Score Clinical Patient Notes**

Identify Key Phrases in Patient Notes from Medical Licensing Exams

PREPARED BY: Jose Baby (Emp. Id. 2211583)

Prepared as part of Kaggle projects given by Mr. Mrityunjoy Pandey.

**About the project:**

The text data presented here is from the USMLE® Step 2 Clinical Skills examination, a medical licensure exam. This exam measures a trainee's ability to recognize pertinent clinical facts during encounters with standardized patients. During this exam, each test taker sees a Standardized Patient, a person trained to portray a clinical case. After interacting with the patient, the test taker documents the relevant facts of the encounter in a patient note. Each patient note is scored by a trained physician who looks for the presence of certain key concepts or features relevant to the case as described in a rubric. The goal of this competition is to develop an automated way of identifying the relevant features within each patient note, with a special focus on the patient history portions of the notes where the information from the interview with the standardized patient is documented

Here, I will be trying to extract important keywords from the medical notes and not do the identification of the relevant features.

**Key feature in the dataset:**

* **Patient Note**: Text detailing important information related by the patient during the encounter (physical exam and interview).

**Steps that I have followed throughout the project:**

1. Installing rake-nltk library which is basically a Rapid Automatic Keyword Extraction algorithm which is domain independent keyword extraction algorithm which tries to determine key phrases in a body of text by analyzing the frequency of word appearance and its co-occuance with other words in a text.
2. Performing a basic text cleaning by removing the punctuations.
3. Creating a dataframe with the extracted key-phrases and corresponding to their scores.
4. Keeping only those key-phrases which have a score greater than 5.
5. Then using Bag of Words and Tf-idf to extract the important key-words.