Medical Abstract Classification

Detailed Description:

Develop predictive models that can determine, given a medical abstract, which of 5 classes it falls in.

Medical abstracts describe the current conditions of a patient. Doctors routinely scan dozens or hundreds of abstracts each day as they do their rounds in a hospital and must quickly pick up on the salient information pointing to the patient's malady. Trying to design assistive technology that can identify, with high precision, the class of problems described in the abstract. In the given dataset, abstracts from 5 different conditions have been included: digestive system diseases, cardiovascular diseases, neoplasms, nervous system diseases, and general pathological conditions.

Since the dataset is imbalanced, the scoring function will be the **F1-score** instead of Accuracy.

Caveats:

The dataset has an imbalanced distribution i.e., there are different numbers of samples in each class.

Data Description:

The training dataset consists of 14442 records and the test dataset consists of 14438 records. The data are provided as text in train.dat and test.dat

train.dat: Training set (class label, followed by a tab separating character and the text of the medical abstract).

test.dat: Testing set (text of medical abstracts in lines, no class label provided). format.dat: A sample submission with 14438 entries randomly chosen to be 1 to 5