Medical Text Classification

In this assignment, we will build a classifier using a medical-NLP corpus.

This is a classification task where the input is a medical transcript (text) and the output is the corresponding medical transcript type.
The dataset consists of transcripts that fall into one of the four classes:
1. Surgery (Surgery - 1)
2. Medical Records (Medical Records - 2)
3. Internal Medicine (Internal Medicine - 3)
4. Other (Other - 4)

The goal is to classify each transcript into the correct class.

Dataset details:
- 4000 transcripts for training
- 500 transcripts for validation
- 500 transcripts for testing

1. Text Vectorization

Since classification algorithms require fixed-length vector input and our texts have variable lengths, we first need to convert the texts into vectors.

In this assignment, we consider two ways of representing the texts:
- Binary Bag-of-Words (BBoW)
- Frequency Bag-of-Words (FBoW)

Both representations must be implemented and prepared.

2. Classification using BBoW

In this section, we use the binary bag-of-words representation and evaluate models using the F1-score.

(a) Train Logistic Regression, Decision Tree, Random Forest, XGBoost.
(d) Report F1-scores for training, validation, and test sets.
(e) Analyze the results and explain the role of hyperparameters.

3. Classification using FBoW

This section is similar to the previous one but uses the frequency bag-of-words representation.

(a) Identify which representation (BBoW or FBoW) performed better and why.

Instructions for BBoW

1. Use only the training set to build the vocabulary (do not use validation or test sets).
2. Preprocess the text: remove punctuation and convert to lowercase.
3. Keep the top 10,000 most frequent words.
4. Represent each text as a 10,000-dimensional vector (1 if the word appears, 0 otherwise).

Instructions for FBoW

1. Follow steps 1 and 2 as above.
2. For each word, count its occurrences in the text, then divide by the total

word count of the text so that the vector sums to 1.

Format of Deliverables

Vocab format: each line contains a word, its ID, and frequency.
Example:
the 1 20456

Train/Valid/Test format: each line is one data point. Words are replaced by IDs, and the class label is placed at the end.
Example:
100 8 3 1034 0

Good luck,
Amir Zamanidoost