Assignment 4 – Machine learning and NLP

Preliminaries

The programming language for this assignment is Python 3. The assignment should be submitted through ilearn no later than the 27th of October at 23.59 (CET).

Decision Tree and Random Forrest:

1. Open the file **`decision_tree.py`**. The class **`BinaryDecisionTree`** implements a binary decision tree classifier. Training of the classifier is implemented in the constructor **`__init__(self, ...)`**, while prediction is done using the method **`predict(self, ...)`**. Get familiar with the code and complete the implementation of the function

```
get_information_gain(self, y:list, y_left:list, y_right:list) -> float
```

2. Open the file **`random_forest.py`**. The class **`BinaryRandomForrest`** implements a random forest classifier based on the **`BinaryDecisionTree`** class. Again, training of the classifier is implemented in the constructor **`__init__(self, ...)`**, while prediction is done using the method **`predict(self, ...)`**. Get familiar with the code and complete the implementation of the function

```
get_sample(self, X:dict) -> dict
```

3. Open the file **`metrics.py`**. It includes several metrics, defined around the confusion matrix, for measuring predictive performance. Implement the following functions:

```
get_false_positives(y_true:list, y_pred:list) -> int
get_true_positives(y_true:list, y_pred:list) -> int
get_false_negatives(y_true:list, y_pred:list) -> int
get_true_negatives(y_true:list, y_pred:list) -> int
get_accuracy(y_true:list, y_pred:list) -> float
get_f1(y_true:list, y_pred:list) -> float
```

4. Finally the file **`run_assignment_4a.py`** trains the above models and prints the metrics. Run it and take a look at the printout. Do you see anything strange in the performance of the models? Write your conclusions in a text document and fix the code that led to the errors you found.

5. Try different values for the model-parameters **`bias`** and **`max_depth`**. How do they influence the outcome? Why? Write your conclusions in the text document.

Bag of Words:

6. Open the file **`run_assignment_4b.py`**. It uses the decision tree for text classification on movie reviews. Manipulate the code for text preprocessing in order to improve the performance of the model. Describe and motivate your alterations and their effects in the text document.

Summary of Assignment:

Your groups assignment should contain the code that you have produced:

- 'decision_tree.py',
- 'metrics.py',
- 'random_forest.py',
- 'run_assignment_4a.py',
- 'run_assignment_4b.py'

All of these should be completed according to the above tasks. Additionally your solution should contain a text or pdf document named 'LastName1_LastName2_ass4' conationing your answers to tasks 4 to 6. Add all files above into a zip file with the last names of both group members, like 'LastName1_LastName2_ass4.zip'