**Assignment 4**

**4. 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.**

In each run we see completely different outcomes, and data is imbalanced. So we would need ti increase the data.

**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.**

If we specify max\_depth = 20 and bias = 0.9 than we see a decrease in accuracy and F1. If we specify max\_depth = 2 and bias = 0.2 than we see again decrease in accuracy and F1. So we need to seek a balance.

**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.**

These are the changes we implemented:   
#1 Lowercasing  
#2 Tokenization and Stop Word Removal  
#3 Stemming or Lemmatization  
#4 Remove punctuation and special characters

First and foremost, it extended the execution time of the Python file due to the data manipulation involved. These steps enhanced the model's performance, which was evident when comparing its performance before and after the incorporation of these NLTK data manipulation techniques. Furthermore, recall is affected badly but precision increased massively.