CS 6375
ASSIGNMENT _____

Names of students in your group:
Sunit Mathew

Number of free late days used:1

_____

Note: You are allowed a total of 4 free late days for the entire semester. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

Part 2

Assumptions Made

1. If there is a node where both the children are leaf nodes of the same type, then the node is merged into one node of that type. This is done recursively, sometimes even higher nodes can be merged because of this same logic.
2. Every time pruning is done, the node to be pruned is chosen randomly from 3 nodes. These 3 nodes are the ones that have the lowest entropy among all nodes.
3. If there is leaf node where the class is not clear, then majority trumps the class.

Accomplishments

1. Decision Tree that works on binary data
2. Pruning in decision tree from a guess

Things Learned

1. How to construct a decision tree
2. How to tweak pruning factor to get higher correlation between test data and training data
3. How to approach over fitting

Log of run:

```
[sunit@mathew ~]$ python Assignment1.py
"/home/sunit/Documents/ClassWork/MachineLearning/Assignments/data/data
/train.dat"
"/home/sunit/Documents/ClassWork/MachineLearning/Assignments/data/data
/test.dat" 0.1



--------------------------------
Pre Pruned Decision Tree
--------------------------------
| tea ID :1 = 0 :
| | honor ID :2 = 0 : 0
| | honor ID :2 = 1 :
| | | romulan ID :4 = 0 : 0
| | | romulan ID :4 = 1 :
| | | | poetry ID :6 = 0 :
| | | | | barclay ID :7 = 0 : 0
| | | | | barclay ID :7 = 1 :
```

| | | | | | wesley ID :9 = 0 : 1
| | | | | | wesley ID :9 = 1 : 0
| | | | poetry ID :6 = 1 :
| | | | | wesley ID :12 = 0 :
| | | | | | barclay ID :13 = 0 : 0
| | | | | | barclay ID :13 = 1 : 1
| | | | | wesley ID :12 = 1 :
| | | | | | barclay ID :16 = 0 : 0
| | | | | | barclay ID :16 = 1 : 1
| tea ID :1 = 1 :
| | honor ID :19 = 0 :
| | | poetry ID :20 = 0 : 0
| | | poetry ID :20 = 1 :
| | | | barclay ID :22 = 0 : 1
| | | | barclay ID :22 = 1 : 0
| | honor ID :19 = 1 :
| | | romulan ID :25 = 0 :
| | | | wesley ID :26 = 0 :
| | | | | barclay ID :27 = 0 : 1
| | | | | barclay ID :27 = 1 :
| | | | | | poetry ID :29 = 0 : 0
| | | | | | poetry ID :29 = 1 : 1
| | | | wesley ID :26 = 1 :
| | | | | poetry ID :32 = 0 : 0
| | | | | poetry ID :32 = 1 :
| | | | | | barclay ID :34 = 0 : 1
| | | | | | barclay ID :34 = 1 : 0
| | | romulan ID :25 = 1 :
| | | | barclay ID :37 = 0 :
| | | | | poetry ID :38 = 0 :
| | | | | | wesley ID :39 = 0 : 1
| | | | | | wesley ID :39 = 1 : 0
| | | | | poetry ID :38 = 1 :
| | | | | | wesley ID :42 = 0 : 1
| | | | | | wesley ID :42 = 1 : 0
| | | | barclay ID :37 = 1 :
| | | | | wesley ID :45 = 0 :
| | | | | | poetry ID :46 = 0 : 1
| | | | | | poetry ID :46 = 1 : 0
| | | | | wesley ID :45 = 1 : 0

Pre Pruned Accuracy

Number of training instances = 800
Number of training attributes = 6
Total number of nodes in the tree = 49
Number of leaf nodes in the tree = 25
Accuracy of the model on the training dataset = 0.8926342072409488
Number of testing instances = 800
Number of testing attributes = 6
Accuracy of the model on the testing dataset = 0.8480392156862745

Number of nodes to be pruned = 4.9
Nodes to be pruned = 45
Nodes to be pruned = 13
Nodes to be pruned = 46
Nodes to be pruned = 2

Post Pruned Decision Tree

| tea ID :1 = 0 : 0
| tea ID :1 = 1 :
| | honor ID :19 = 0 :
| | | poetry ID :20 = 0 : 0
| | | poetry ID :20 = 1 :
| | | | barclay ID :22 = 0 : 1
| | | | barclay ID :22 = 1 : 0
| | honor ID :19 = 1 :
| | | romulan ID :25 = 0 :
| | | | wesley ID :26 = 0 :
| | | | | barclay ID :27 = 0 : 1
| | | | | barclay ID :27 = 1 :
| | | | | | poetry ID :29 = 0 : 0
| | | | | | poetry ID :29 = 1 : 1
| | | | wesley ID :26 = 1 :
| | | | | poetry ID :32 = 0 : 0
| | | | | poetry ID :32 = 1 :
| | | | | | barclay ID :34 = 0 : 1
| | | | | | barclay ID :34 = 1 : 0
| | | romulan ID :25 = 1 :
| | | | barclay ID :37 = 0 :
| | | | | poetry ID :38 = 0 :
| | | | | | wesley ID :39 = 0 : 1
| | | | | | wesley ID :39 = 1 : 0
| | | | | poetry ID :38 = 1 :
| | | | | | wesley ID :42 = 0 : 1
| | | | | | wesley ID :42 = 1 : 0

| | | | barclay ID :37 = 1 : 0

Post Pruned Accuracy

Number of training instances = 800
Number of training attributes = 6
Total number of nodes in the tree = 29
Number of leaf nodes in the tree = 15
Accuracy of the model on the training dataset = 0.8626716604244694
Number of testing instances = 800
Number of testing attributes = 6
Accuracy of the model on the testing dataset = 0.8137254901960784