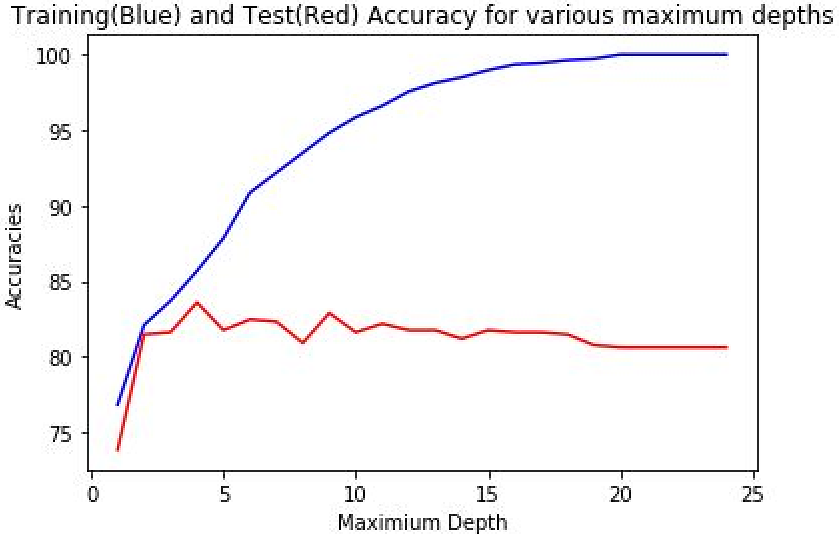
(a) Graph



(b) Overfitting occurs as test accuracy starts to decrease after reaching a peak while training accuracy keeps on increasing

Overfitting starts to occur after DEPTH 4 (test accuracy declines after reaching a peak at depth 4)

(c) The tree that achieved the highest testing accuracy was the tree with depth 4.

---Decision Tree using Information Gain--- Depth 4

Note: Over here the number (let’s say 484 represents the word on line 484+1 = 485 in the words.txt file)

```
484=0

|211=0

     |152=0

          |73=0:comp.graphics

          |73=1:alt.atheism

     |152=1

          |187=0:comp.graphics
```

```
        |187=1:alt.atheism
|211=1
    |183=0:alt.atheism
    |183=1
        |0=0:comp.graphics
        |0=1:alt.atheism
484=1
|3142=0
    |2108=0
        |152=0:alt.atheism
        |152=1:alt.atheism
    |2108=1:comp.graphics
|3142=1:comp.graphics
```

We can see that at each node the tree has decided to gain the maximum information it could at that stage. This is the greedy approach and works to optimise the tree in an optimal time. Hence, the features chosen by our model makes sense and it is the best in terms of generalising with other data i.e. test data as it does not grow too big to start overfitting the training data. Hence, all the features used made sense.