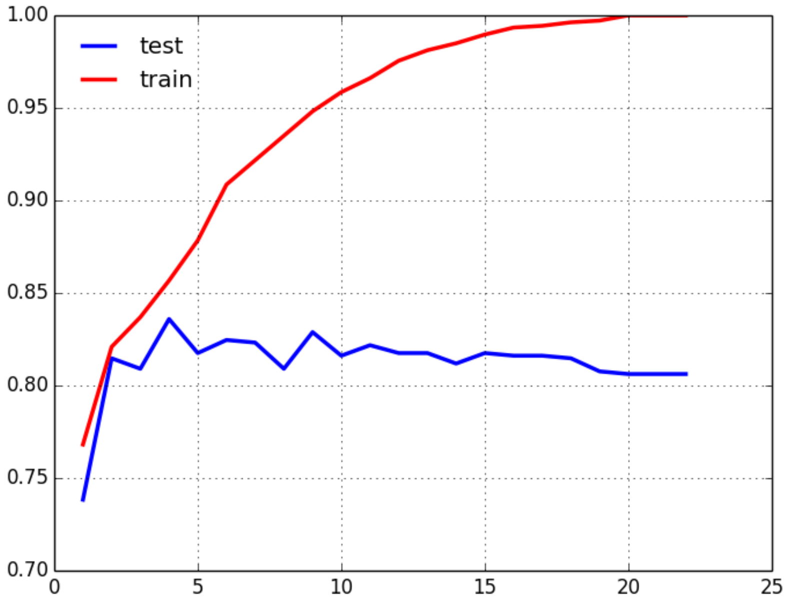
CS686 – assignment 3

Name: Erlie Shang

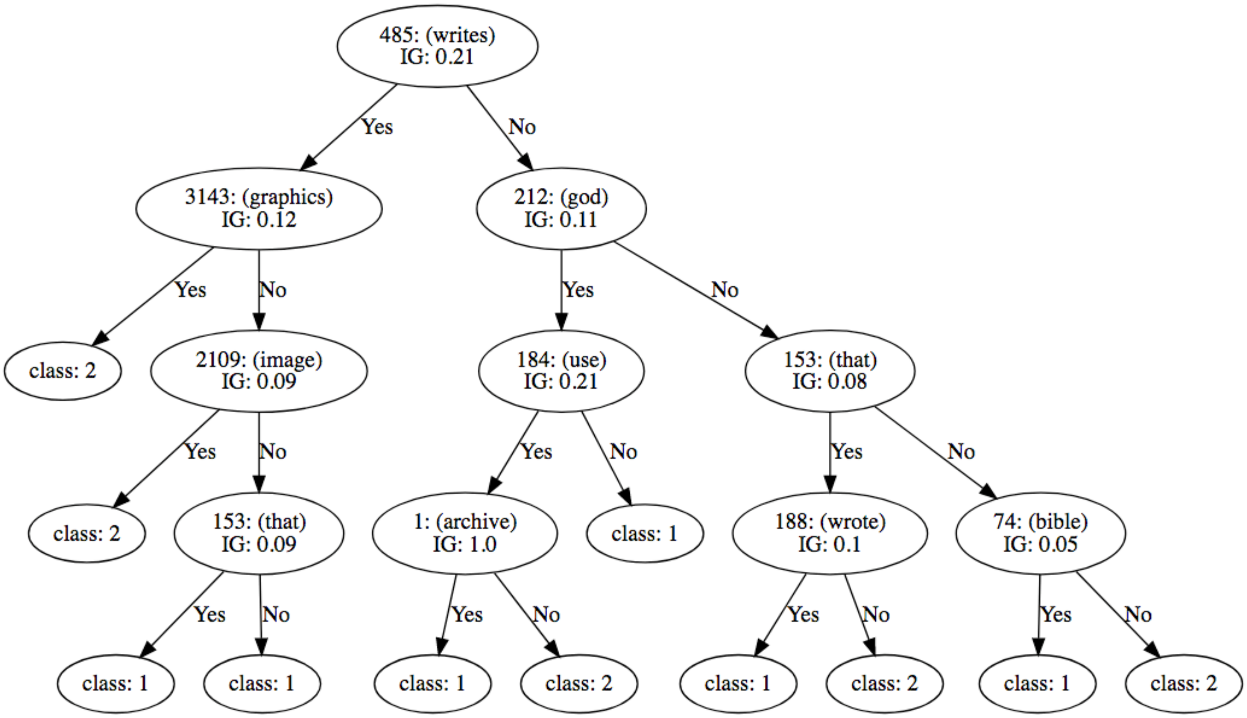
Student#: 20696176

1. Decision Tree Learning
   1) See DTL.py. You will need **graphviz** and **matplotlib** to run the program. After running this program, you will see **accuracy.png** and **deicisionTree.svg** in the working directory, which illustrates the accuracy rate and the decision tree that achieved the highest testing accuracy.
   2)



Note: The depth of this tree is defined as the numbers of the edges in the longest branch.

   3) Yes. It happens after the maximum depth of 4.
   4)



This is the tree that achieved the highest testing accuracy rate. For each node, the number is the index of the word that related to it, and the IG denotes the information gain. Yes on the edges means that the data contains this word, while no means not. Class 1 is alt.atheism and class 2 means comp.graphics.

5) In my opinion, not all the word features selected make sense. Words like graphics and image are more likely to appear in the articles from comp.graphics, while words like god and bible might show up in the articles from alt.atheism more frequently. However, there are also some words that don’t make sense. For example, the words like writes and that are really common in our daily writing, so they are unlikely to be the features of one particular kind of articles.

2. Naive Bayes Model

1) See MLL.py. After running the program, you will see the accuracy rate and the word list on the terminal.
2)
```
(4.4245173392155825, 'graphics')
(3.974792811543938, 'atheism')
(3.9282727959090455, 'religion')
(3.8541648237553234, 'moral')
(3.8541648237553234, 'keith')
(3.8541648237553234, 'evidence')
(3.828189337352063, 'atheists')
(3.7833387711867106, 'god')
(3.7459512391150906, 'bible')
(3.7169637022418387, 'religious')
```
   In my opinion, not all these words are good word features, but most of them are. Graphics might mean the comp.graphics, while words like atheism, religion, moral and so on are more relevant to alt.atheism. However, words like keith seems to be irrelevant.
3) The training accuracy is: 0.928369462771
   The testing accuracy is: 0.889674681754
4) No, of course not. An article is likely to contain bible and god at the same time.
5) An edge between independent features can be added.
6) The naïve Bayes model performed better. The reason is that the naïve bayes model does not have overfitting, while the decision tree starts to overfit quickly. Besides, the words selected by Naïve bayes model seem to be more reasonable.