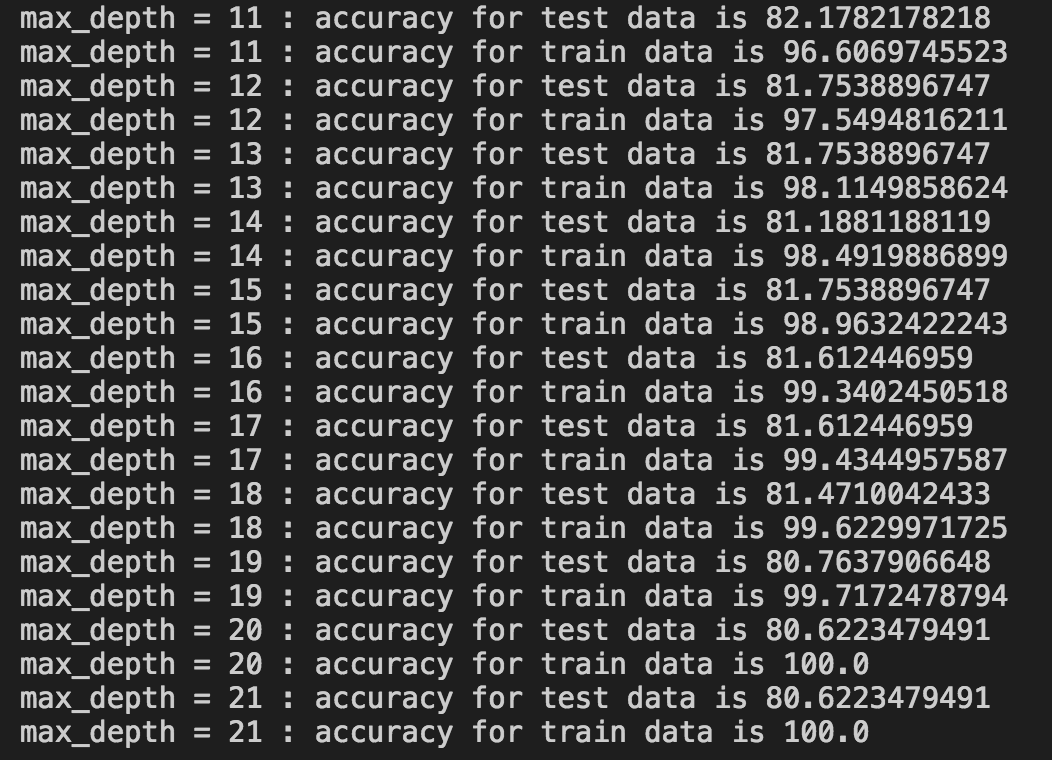
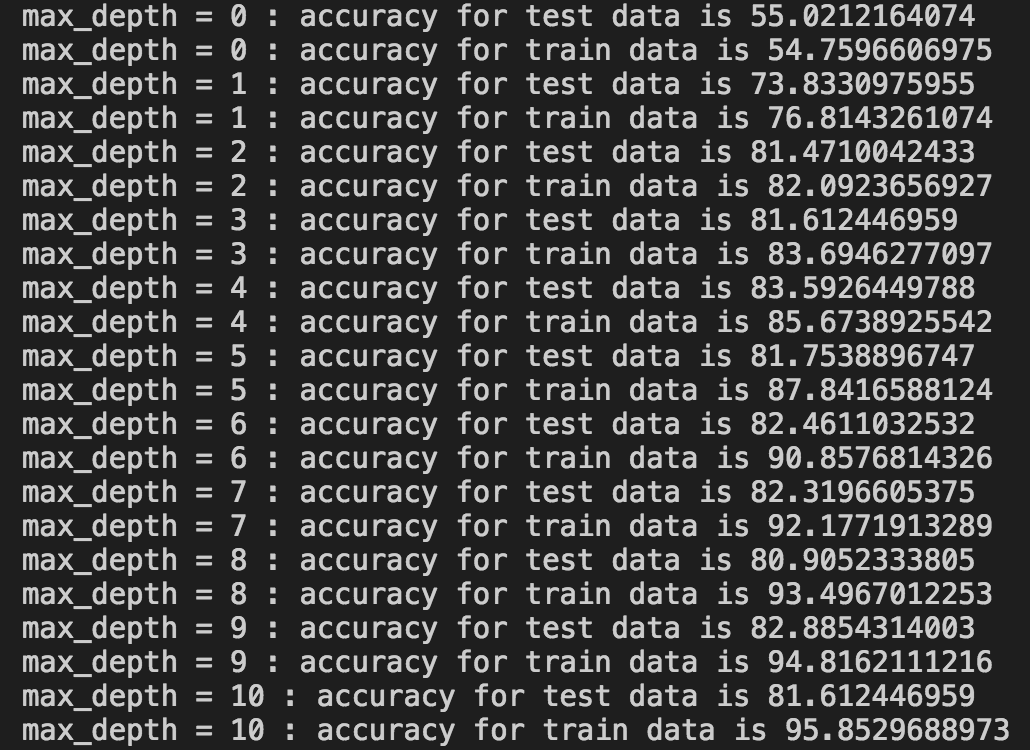
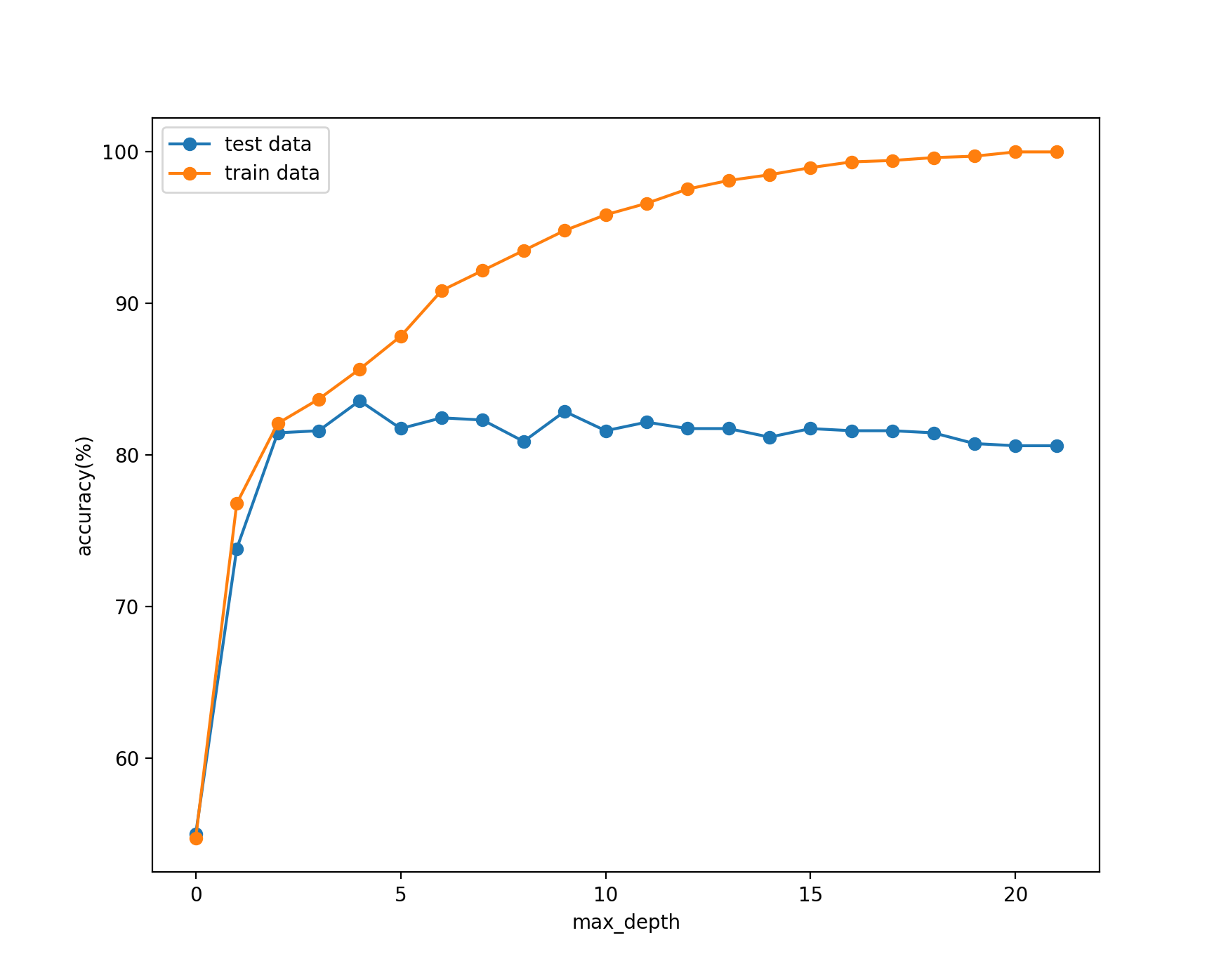
Ruixue Zhang 20619404

Problem 1. Decision Tree Learning

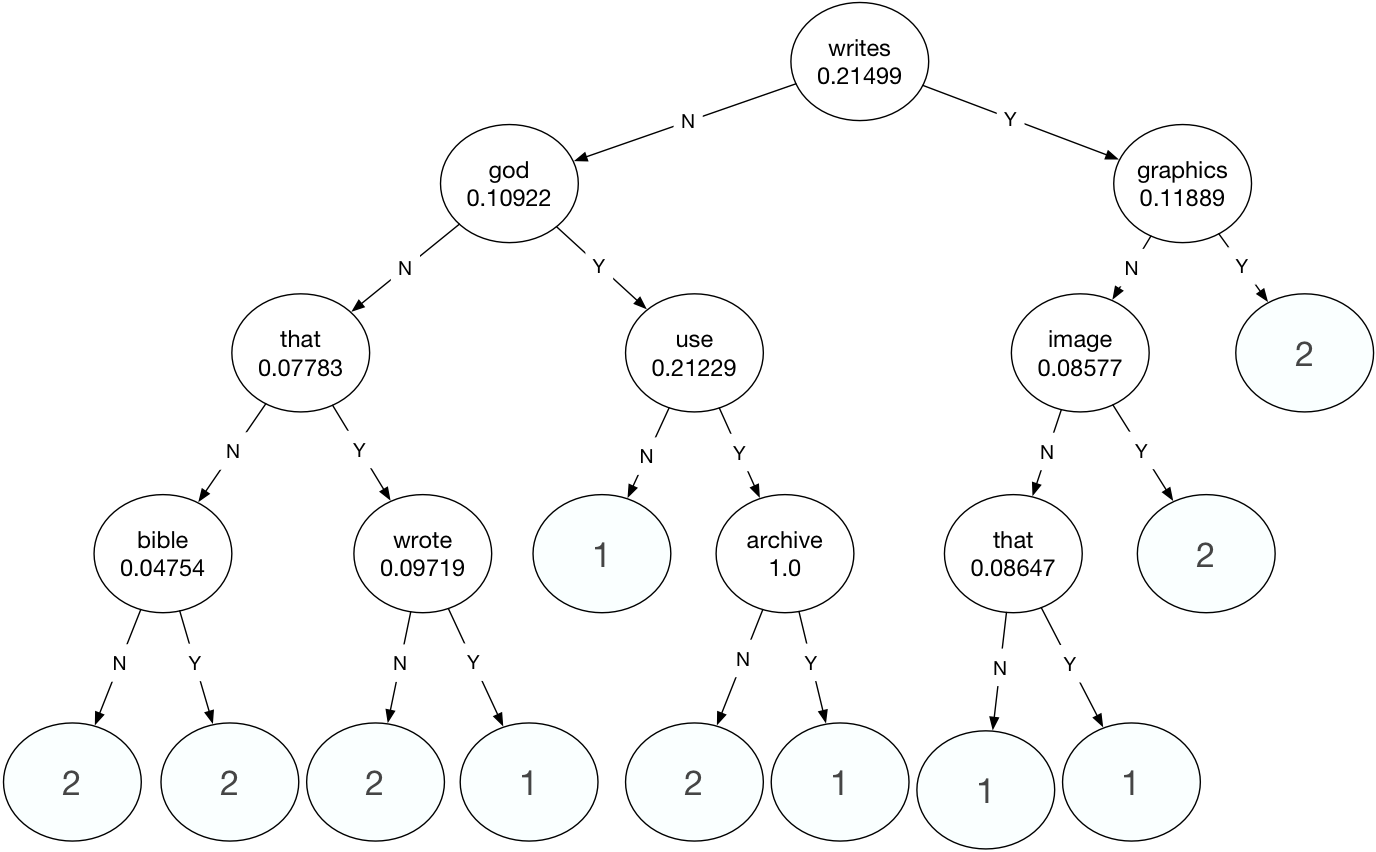
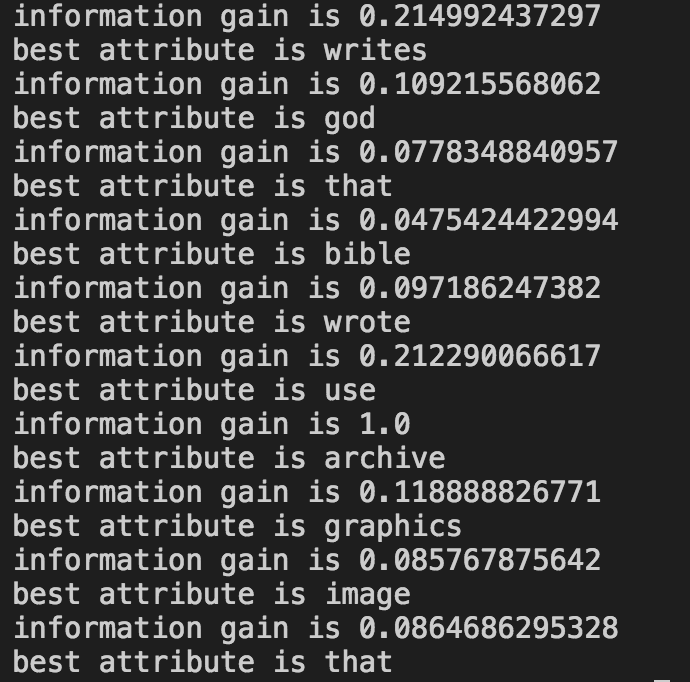
1. The code is attached in the folder.

2. Training and testing accuracy as the maximum depth increases:



3. Yes. Overfitting occurs after maximum depth equals to four.

4. The decision tree that achieved the highest testing accuracy (83.6%):

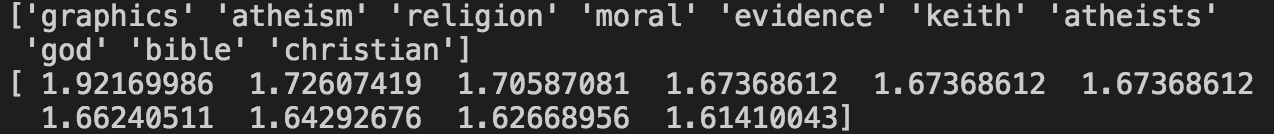


5. Some words selected do not make sense. Words like ‘that’ will appear many times in an article and have no special meanings. This won’t help to do text categorization.

Problem 2. Naïve Bayes Model

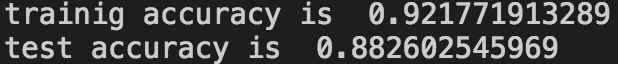
1. The code is attached in the folder.

2. 10 most discriminative word features:



These words are good to discriminate the two categories. Graphics is evident to category 2 (comp.graphics). Atheism, religion, moral, atheists, god, bible and Christian are highly related to category 1 (alt.atheism). Evidence seems to belong to category 1 since we more likely to mention evidence when we talk about something about religion. Keith is a film in google search result, which should a good word to find class comp.graphics.

3. training and testing accuracy:



4. This is not a reasonable assumption since one word may be different part of speech and language exists polysemy phenomenon. This should be justified based on context.

5. In Naïve Bayes model, we assume that all words are independent, then there is no linkage between different features. To extend Naïve Bayes model, we can link different features if these words highly correlated to build a Bayesian Network.

6. Naive Bayes model (88.3%) performs better than decision tree learning (83.6%). Decision tree learning chooses features based on information gain in this case. Some features may have almost same information gain but the ‘best’ one selected may not be very reasonable according to the last problem. When the tree is applied to the test set, accuracy won’t be very high.