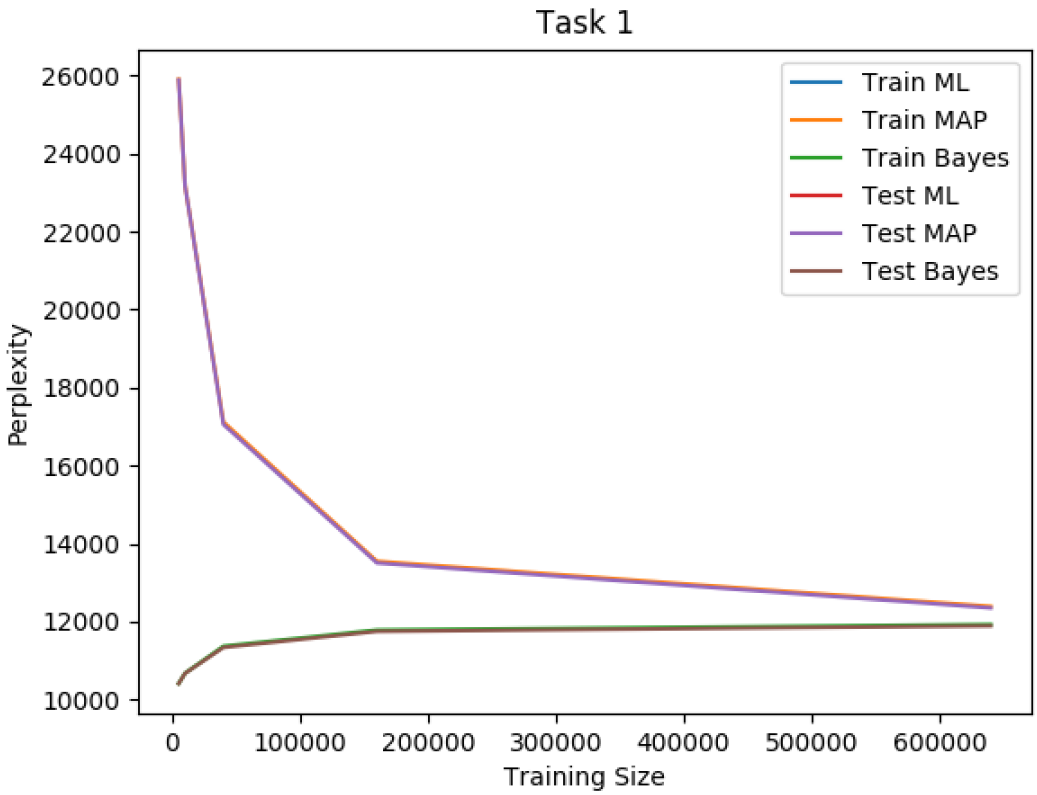
Christian Zinck

Programming Project 1

Task 1:

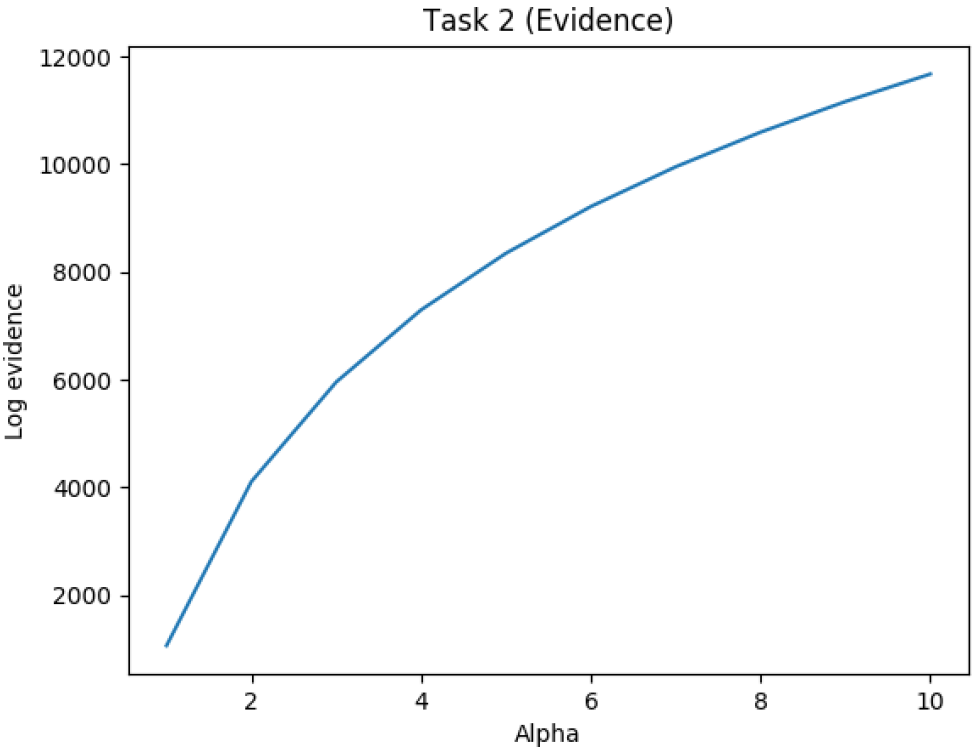


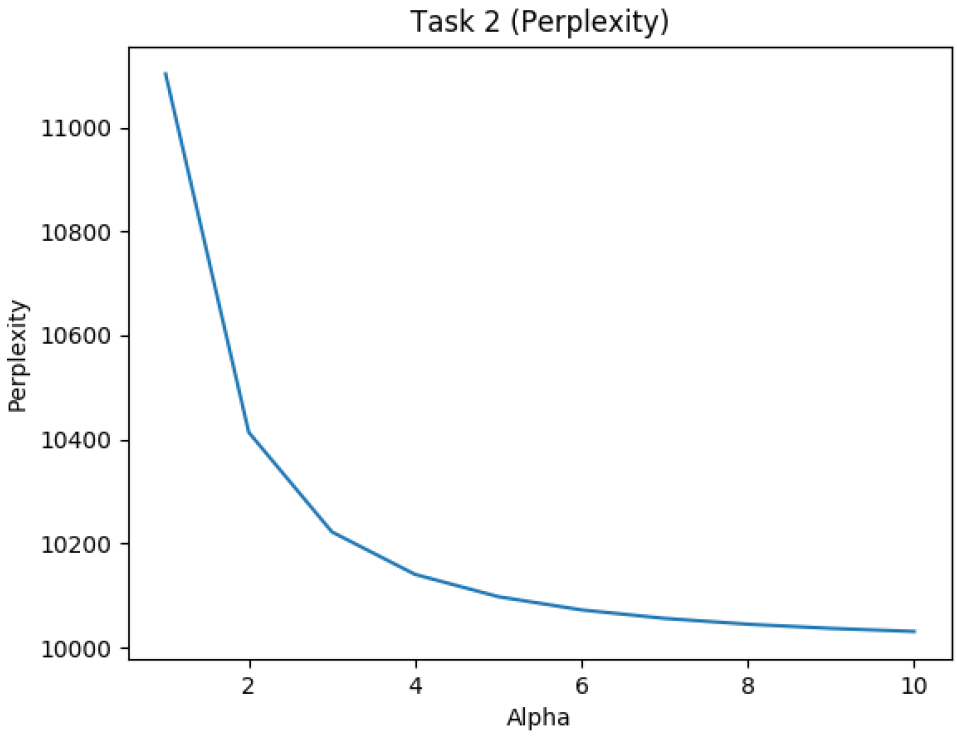
As can be seen in the graph above, as training size increases, the perplexity for both test and training sets converges. The test set begins at a high perplexity and decreases, because as more information is gained from a larger training set, the test set is better predicted. The training set begins at low perplexity and increases, because it is easier to predict the contents of a small set compared to a larger one. In either case, when the full training set is used, the test set and training set have similar perplexities.

It may not be obvious from this graph, but the maximum likelihood estimation for the test and training sets produced an infinite perplexity for all training set sizes other than the complete set. This is because if one a word is missing from the training set, the maximum likelihood model will predict that it is impossible for that word to occur, which is a very bad prediction. The MAP and Bayes estimates perform better on incomplete data sets because they make use of regularizations.

With the full training set, the perplexity of the ML estimate is very similar to the other training sets, which does not make use of regularization at all. Therefore, the perplexity of the test set should not be very sensitive to a small change in alpha for the full training set.

Task 2:


Task 2 (Evidence)
Log evidence
12000
10000
8000
6000
4000
2000
2
4
6
8
10
Alpha


Task 2 (Perplexity)
Perplexity
11000
10800
10600
10400
10200
10000
2
4
6
8
10
Alpha

From the graphs above, we can see that as alpha increases, evidence increases and perplexity decreases. Since a high evidence and a high alpha results in a low perplexity, which is good, maximizing evidence is a good method for model selection on this data set.

Task 3:
Using page 345 to train, alpha = 2 and the bayes prediction model, the perplexity of page 1188 was 27490.3557922 while page 84 was 28148.4046582. The model correctly predicted that page 1188 was by the same author as page 345, but the margin by which the prediction was made is by no means impressive. This is most likely due to the fact that all 3 novels are horror/mystery novels written in a similar time period in Europe. I would expect a set of novels with more word variety to perform better under this model.