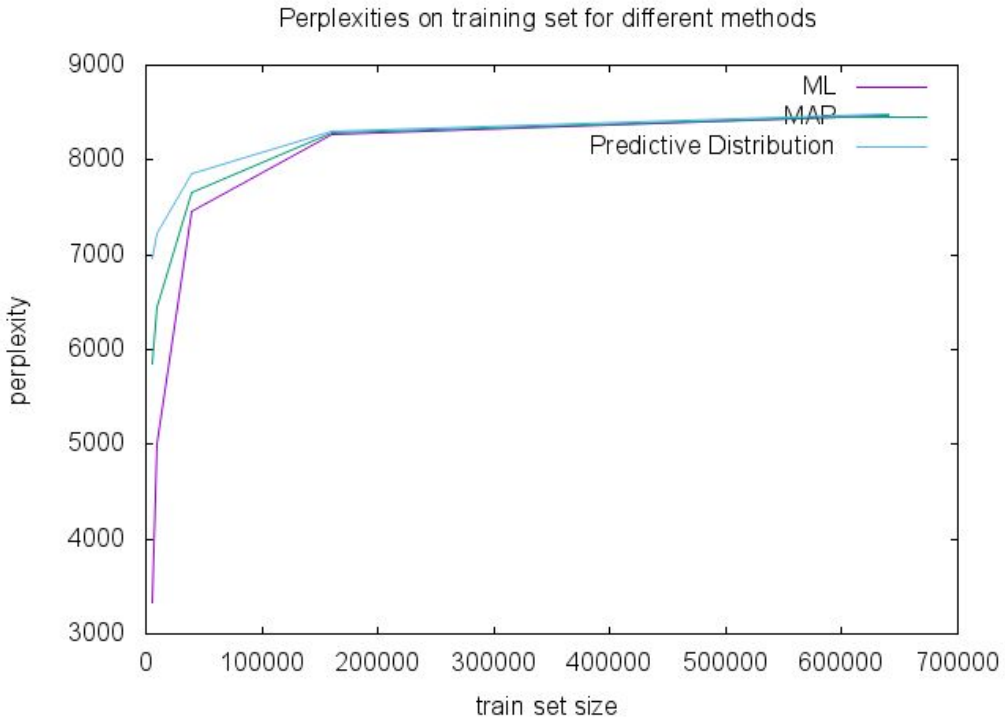
Report for Project 1

Zhaokun Xue

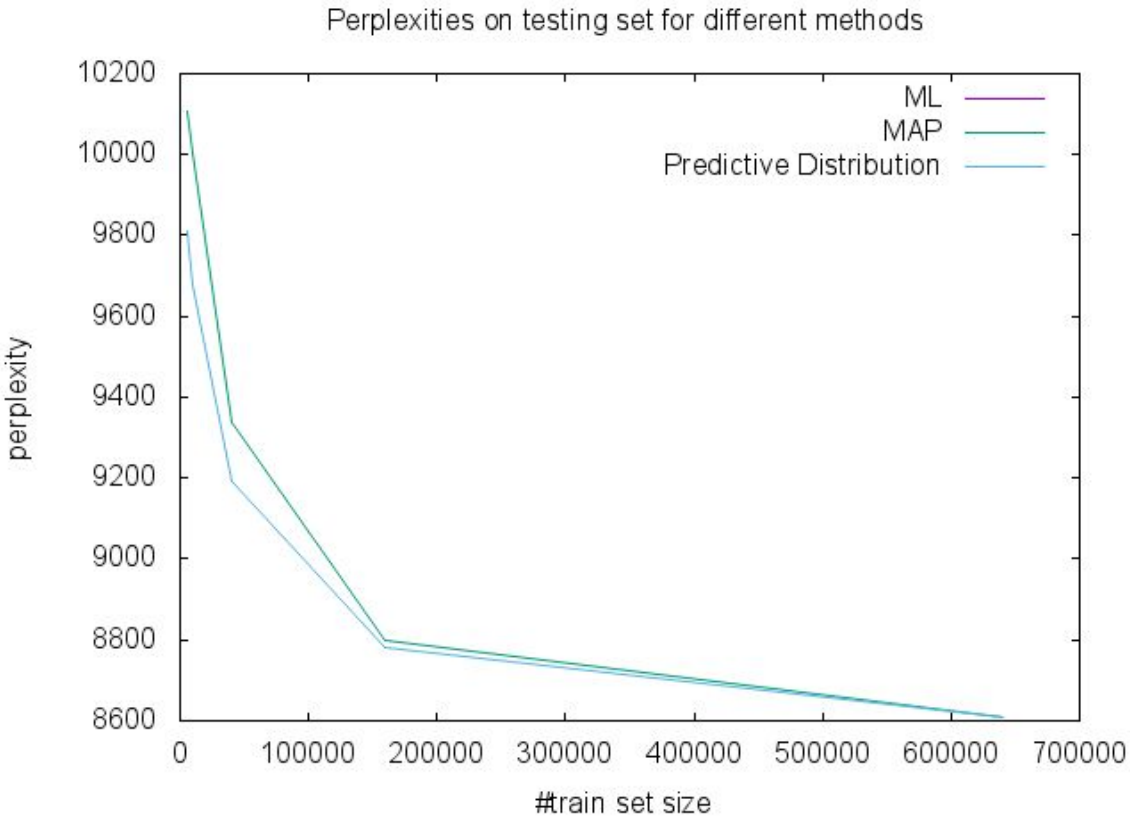
- **Task1**
  - **Perplexity for train set**

Train set size	ML	MAP	Predictive Distribution
5000	3337.69	5850.78	6953.93
10000	5010.34	6452.85	7226.36
40000	7462.09	7653.89	7851.68
160000	8276.40	8286.92	8308.04
640000	8476.45	8477.00	8478.50



- **Perplexity for test set**

Train set size	ML	MAP	Predictive Distribution
5000	Infinity	10106.88	9812.38
10000	Infinity	10004.36	9677.34
40000	Infinity	9338.60	9191.11
160000	Infinity	8800.54	8779.92
640000	8612.35	8609.54	8607.97



Note: There would only be one point(640000, 8612.35) for ML on this graph. All the other values for ML would be infinity, which cannot be plotted by gnuplot.

- **Discussion**

1. **What happens to the test set perplexities of the different methods with respect to each other as the training set size increases? Please explain why this occurs.**

   For all three methods, the test set perplexities decrease as the training set size increasing. As the training set size increases, we will have less unknown words, which means we have less words with small probabilities or even with probability 0. This will give us a higher result from ln(p(w_i)) which will lead to a lower perplexity.

2. **What is the obvious shortcoming of the maximum likelihood estimate for a unigram model? How do the other two approaches address this issue.**

   The maximum likelihood estimate causes overfitting when training with incomplete training set. When a word appears in test set but not training set, its ML would be 0. And when we calculate perplexity, the result would be infinity as we have ln(0) in our calculation.

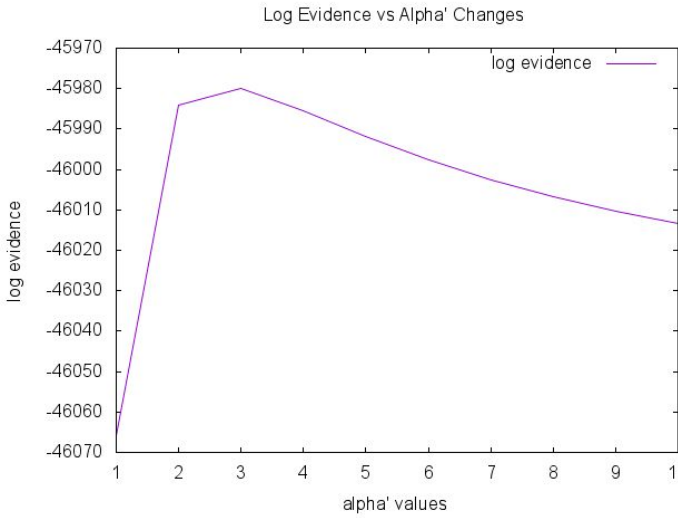
   The other two approaches address this issue by adding a smoothing term *p* to the numerator and a smoothing term *v* to the denominator, where *v* = *kp*. For MAP, *p* = α_k - 1, *v* = α_0 - K and k = K. In predictive distribution, *p* = α_k, *v* = α_0, and k = K. This approach makes sure we never get 0 for probabilities. And we can control this smoothing term by changing the value of α' in our case.

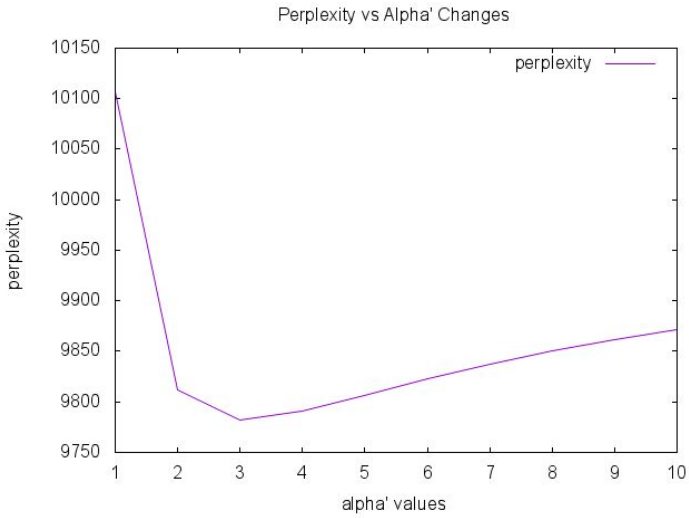
3. **For the full training set, how sensitive do you think the test set perplexity will be to small changes in α' ?**

First, The test set perplexities for ML will not change, as it does not depend on α'. As for MAP and Prediction Distribution methods, their perplexities will not be much sensitive to small changes in α'. Although the changes in α' will affect the value α_k in the numerator, this change will be largely smoothed by the change of α_0 in the denominator, where α_0 = K*α_k. This makes the effect of small changes in α' really trivial. Therefore, the perplexities for these two methods will not be much sensitive to small changes in α'.

- **Task 2**

alpha'	Log evidence	Perplexity
1.0	-46065.89	10106.88
2.0	-45984.14	9812.38
3.0	-45979.97	9781.82
4.0	-45985.38	9790.74
5.0	-45991.76	9806.79
6.0	-45997.66	9822.95
7.0	-46002.52	9837.57
8.0	-46006.75	9850.41
9.0	-46010.36	9861.59
10.0	-46013.46	9871.35





- **Discussion**

1. **Is maximizing the evidence function a good method for model selection on this dataset?**

   Based on our implementation, we can conclude that at α’ = 3, we have the smallest perplexity and we get the max evidence. Therefore, maximizing the evidence function is a good method for model selection on this dataset.

● Task 3

Perplexity for pg84	Perplexity for pg1188
8270.715156571314	5864.369456955972

According to our calculation, we can say that pg1188 has the same author as our training file pg345. The model was successful in classifying these two testing files. Because there is a significant difference between the perplexities of these two testing files. We can feel confident to say that the one with lower perplexity has the same author as our training file.