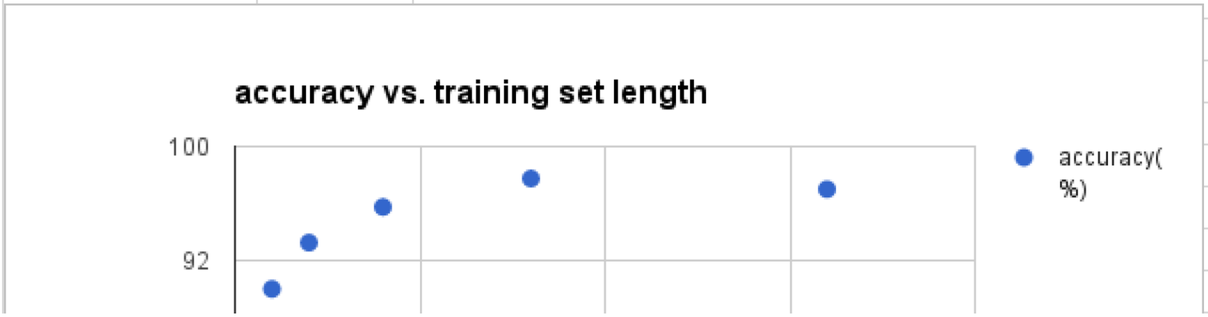
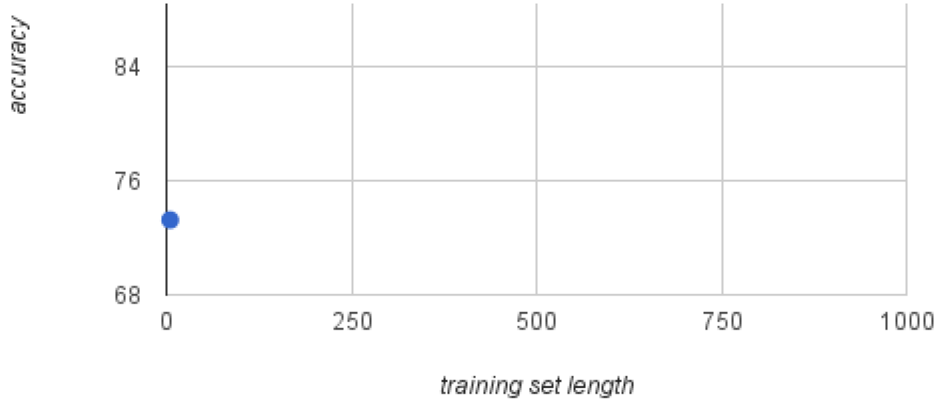
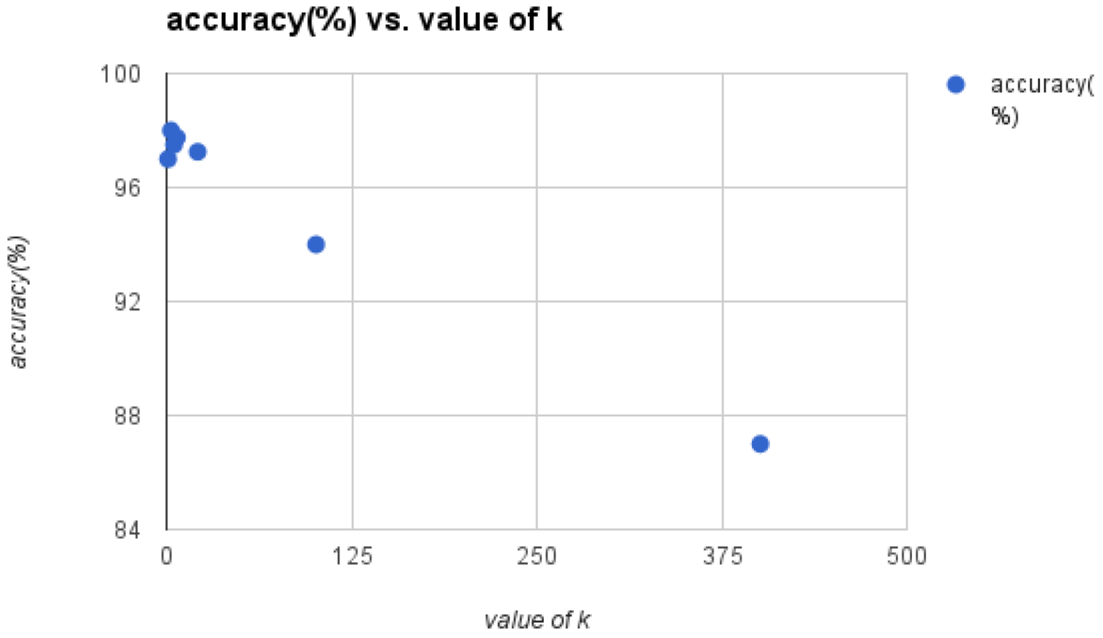
test size of training data with k=1		
training set length	accuracy(%)	
5	73.25	
50	90	
100	93.25	
200	95.75	
400	97.75	
800	97	
		conclusion: accuracy improves with more training data. Possible explanation is that more available neighbours provides closer neighbours, thereby increasing accuracy that a test item is equal to its neighbours. Graph shows a logarithmic relation between training size and accuracy
Overfitting and Underfitting with N=100		
value of k	accuracy(%)	
1	97	
3	98	
5	97.5	
7	97.75	
21	97.25	
101	94	
401	87	
		Accuracy peaks at around k values between range 3 to 7. Underfitting will increase chance of wrong estimate due to too small a sample size, and overfitting will capture too wide a group, that will increase chance of including error data. I believe the exact k value of optimal accuracy will depend on the nature of the data which determines the distances, and the training size (grows in proportion to the training size). It would be possible to create a formula to optimize value of k that is a function of training size and another factor for nature of data




accuracy
84
76
68
0
250
500
750
1000
training set length


accuracy(%) vs. value of k
100
96
92
88
84
accuracy(%)
0
125
250
375
500
value of k
● accuracy(%)