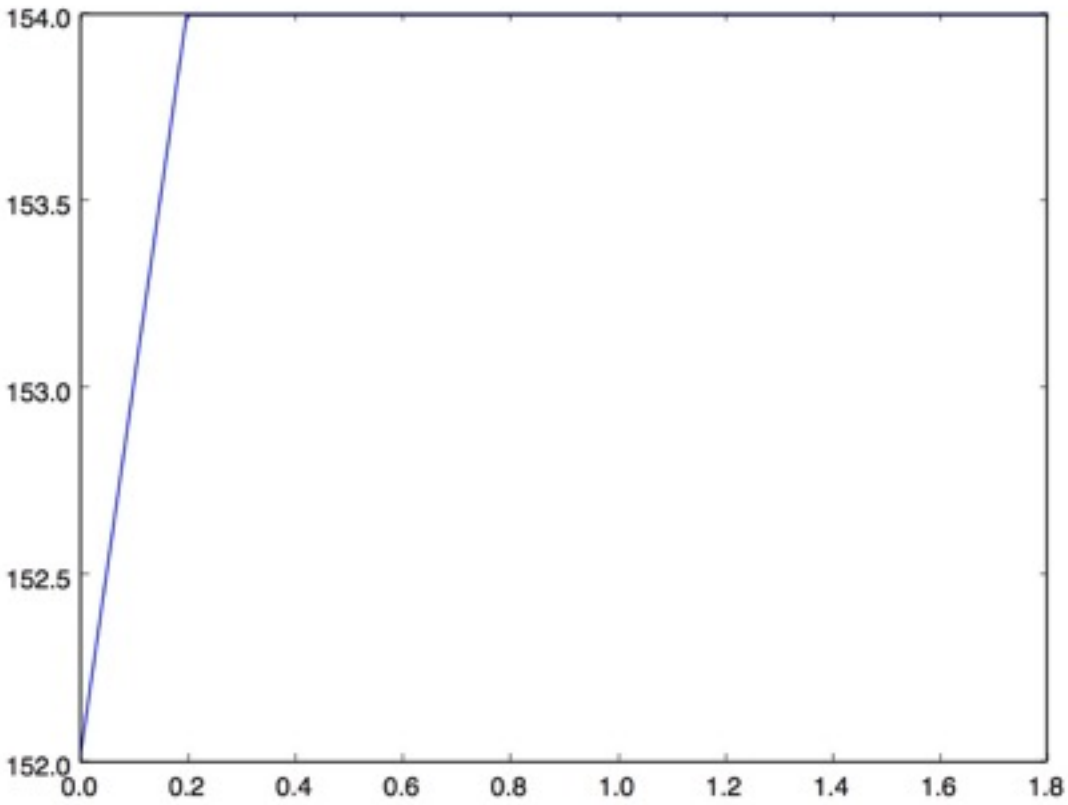
HW 8 Q2

Solution to 4th part of Q2:

a. We have considered twitter ID, Sentiment Score and Twitter Text as the features of the our project data set. We are currently analyzing the data set for one particular state : Arkansas, so we have not take state as a feature but sooner we will be taking the state as the feature.
b. The trade off parameter C of SVM and logistic regression was identified by incrementing C from 0 to 2 with step size of 0.2. we observed that at a particular C=0.9 we observed that number of positive tweets obtained from SVM become constant and so with the logistic regression. So, C is calculated by incremental process. please see the graph obtained at last of this document. **The x axis of the graph is the C values and the y axis represent the number of positive tweets obtained from SVM.**
c. Regarding the model selection: on executing the SVM and Logistic regression on the tweets we observed that SVM gave 67 % accuracy and Logistic regression gave 66.6 % accuracy. So we conclude that SVM is a more desirable classier for our project.
d. The feature that plays the critical role in our project is the sentiment score obtained using pattern.en module. For this reason we have taken sentiment score of the tweets as the class label in the project as both SVM and Logistic regression perform supervised classification.
e. yes, we can predict top K(10) tweets with high confidence factor using the inherent decision function of SVM and logistic regression. The decision function in SVM is the curve fitting linear equation wx+b=0. Since we are taken 10 maximum tweets at a particular C and we are taking range of C from 0 to 2 with step size of 0.2 so we have 10 tweets in for each C values so in all 100 maximum tweets. These tweets are store in maxtweetsfile.txt in the folder.
f. yes, we can predict top K(10) tweets with less confidence factor using the inherent decision function of SVM and logistic regression. The decision function in SVM is the curve fitting linear equation wx+b=0 . Since we are taken 10 maximum tweets at a particular C and we are taking range of C from 0 to 2 with step size of 0.2 so we have 10 tweets in for each C values so in all 100 maximum tweets. These tweets are store in maxtweetsfile.txt in the folder.
g. We have stored top 10 tweets with high confidence factor obtained from Logistic regression classier in “maxtweetsfileLR.txt” and top 10 tweets with low confidence factor in “mintweetsfileLR.txt”.


154.0
153.5
153.0
152.5
152.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8