Part 1 : Digit Recognition

a) Status and stopping point-

In the first dataset for optical recognition of digits, the maximum accuracy percentage is around ~96% which can be considered as an acceptable value. SVC Classifier is used in this case. Based on the varying values for C, the accuracy varies. The best accuracy is achieved for value of C as 0.001. Also the variation is not that bad. Values less than C=0.00001 can be considered as stopping point. For values lesser than C=0.00001, accuracy drops to 20%.

b) Additional functions and analysis:

Based on analysis done for this data set, I have observed that if we plot graph of accuracy vs value of C, then its a step graph. For values higher than 0.001 the accuracy is 96%, after that it drop to 95% till 0.00001 and then it drops to 20%

c) The most challenging part is the data formation. Understanding and Importing the data and the proper use of libraries took time. So code must be efficient to run over those data. That is what I learnt; to handle data and manipulating the data using efficient code methodologies. Also deciding on value of C is a challenge.

Part 2: Amazon Data Set

a) Status and stopping point -

With Amazon Data Set, the maximum accuracy percentage that I get is around ~58%. Which is less compared to other classifiers used so far. Even after changing values C and gamma dramatically. This classifier is slower and accuracy wasn't that great either. There was no stopping point as such in this case.

b) Additional functions and analysis -

After analysing data I realized that it's necessary to clean the data as data set had rows with no reviews. To get rid of those rows I used dropna() method on data frame.
Then review itself had some html markup words. To counter that I used Beautiful soup. And then in order to get rid of stop words such as 'is','and' I used nltk package.
If I don't perform all these operations the accuracy of the code is lower than 54%.

c) Challenges -

Amazon data set is really huge and playing with such huge data was challenging for me. Also figuring out suitable values for C and gamma is a challenge, as various combinations give almost similar accuracy.