HW8 Question 2

a)

There were few features considered: Firstly, I have used my project tweets, which are related to presidential election. Therefore almost the entire data set is talking about presidential candidates.

And apart from that a dictionary is created which only has keywords with frequency 20 or more, all the other stop words are ignored and are not used to build the prediction model.

b)

The trade of parameter: For the value of C, I tried changing it to get better/different output. In the training data set I have 310 positive tweets from 1103 tweets. For C=1, I got 73 positive tweets from the SVM prediction model and for C=5, I got 79 which is more accurate to the original output.

c)

I have chosen SVM as my model, since it is giving me more accurate prediction model than LR for this data set.

d)

Critical features would be most frequent keywords (For my data set, it would be tweets talking about various candidates) and the other feature would be class labels of the tweet. Every tweet is labelled as 1 for positive and 0 for negative using sentiment values.

e)

Top K tweets: I used clf.decision_function to find top K tweets.

clf.decision_function returns the value of a particular tweet between -1.9 to 3.2 .

For my data set, I have chosen the tweets with value greater than equal to 1 to be my top K tweets.

Therefor there are 16 top tweets.

f)

The same way above explained,

I used clf.decision_function to find the values.

To find the lowest tweets I decided the value less than -1.5 will be the tweets with lowest confidence.

There are 24 tweets with lowest confidence.