Question 2:

Design Decisions:

Note : Parallelly have been working on Sentiment Analysis project for another course. Hence, all the basic steps except POS tagging, Stemming are followed.

- We start with removing line breaks and converting tweets into single line string. To do this we stored 12 possible labels in a list and all lines that do not start with this label are appended to previous line.
- Next, we removed punctuations as they mostly do not have any significance related to location.
- Stopwords like 'the' , 'and' etc. are removed as they do not contribute to accuracy. Having stop words as the list of features did not improve the accuracy.
- In next step we calculate number of times each word occurs and select the most occurring (top 1800) words as the best features.
- Decision for selecting top 1800 words is based on the fact that maximum accuracy is reached at 1800.
- To tackle issue of new words occurring in the test data , we completely ignore those words instead of using the concept of pseudo code. Using Pseudo count did not work for our implementation as it reduced accuracy significantly.
- The maximum accuracy reached for the given set of training and test data is 57.8%.

Note : Given how some tweets in the training data span across multiple lines, we used a stored list of cities and appended all lines to the previous tweet , that do not start with the stored labels. Hence, the implementation only works for the given 12 labels in the training data.