Task 3.1:

Data Preprocessing: It is important to clean the data before applying any kind of machine learning algorithm on it. It helps us to improve the quality of the data thereby assuring accuracy and completeness. The raw data received comprises of a large amount of textual data-messages, emails, books etc. Hence first and foremost will be data cleaning that is removing any kinds of repetition of information or incomplete data or missing data. I would also try to solve any kinds of errors that maybe present so that I don’t have to worry about it later. Next I will try to save it in one big file which contains all the data at one place and will be easy to access. Now as the data is present in different formats such as messages or emails the next step would be to standardize everything and store it in a single format that is maybe only text. Once this is done now it will be easier to dive deep and examine the data carefully. Therefore we filter out all the information that is irrelevant to us and only keep the precise and important data. This will greatly reduce the size of the data and also help us to access it more efficiently. Another optional next step would be to divide the data into categories on the basis of what is the purpose of the dataset and how is it going to be used. This is also called as clustering and helps us to store similar data at a single place thereby making our work easy.

Task 3.2: