Task 3.1:

Data Cleaning : 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 we need to perform data cleaning that is removing any kinds of repetition of information or incomplete data or missing data. It will also be better to sort out errors in the program.

Data Transformation: As the data has been gathered from different sources we will now store it in one single place so that it becomes easier to access and manipulate at once. 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 according to whatever purpose the data is going to serve. For emails and messages it would be better to first convert it into normal text containing only the information which can be inferred from the conversations. Once this is done now it will be easier to dive deep and examine the data carefully.

Data Reduction: We filter out all the information that is irrelevant to us and only keep the precise and important data. As we have data which is even taken from books it will contain a lot of data that is not relevant to us and hence data reduction is important so as to save time. It will greatly reduce the size of the data and also help us to access it more efficiently.

Data Discretization: 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. In this way we can have all the related data in one place.

Task 3.2:

Machine learning can be divided into two parts supervised and unsupervised learning. Unsupervised Learning in AI is a type of machine learning wherein there is no human supervision. The machine learns from the data provided by itself. Mostly for unsupervised machine learning models raw data is fed into it and the machine itself tries to find the patterns without any instructions from a human coder.

One of the techniques used is clustering. Clustering refers to grouping up of data which is similar to each other. Clustering often graphs the data based on different basis and then puts together the data points that are closest to each other. Clustering can be of many types including Centroid based Clustering , Density Based Clustering, Connectivity based Clustering and Distribution Based Clustering.

Centroid Based Clustering also known as Partitioning Methods is mostly used for datasets where the data can be divided into separate defined clusters. It is also used where the number of clusters can be easily determined or estimated. Although this can also prove to be a drawback if we are not able to estimate the number of clusters or they are not well defined. The centroid of a cluster is the mean of all the values contained in that particular cluster. K-means and K-medoids clustering are some examples of this type of clustering.

Density based clustering as the name suggests makes groups on the basis of density of the data points. Hence depending on the density of data points it can automatically by itself form clusters or calculate the number of clusters which was initially a drawback in the previous centroid Based clustering. It also solves the other drawback that is concerned with the overlapping of clusters. Hence it is better to use density based clustering for clusters with a lot of ambiguous data. A popular example for density based clustering would be DBSCAN.