Task 3.1:

Text Preprocessing can be done using NLP. NLP refers to Natural Language Processing which is basically a field that comes under computer science and artificial intelligence(A.I.) which uses machine learning to enables computers to recognize and communicate with human language. There are various phases involved in NLP like phonological , morphological ,lexical, syntactic , semantic and pragmatic analysis. However with concern to text processing the main tasks to be carried out are tokenization, stemming and stop word removal.

Text Cleaning: Lowercasing is done so that the algorithm doesn’t treat the same words differently at different locations. It involves converting all the words in the text to lowercase and is done using text.lower(). Further punctuation marks, extra whitespaces and special characters are also removed to focus on the content itself. Furthermore when we copy data from websites it also comes with unnecessary URLs or HTML tags. Hence its important to remove these things as well.

Tokenization : Word tokens are the most fundamental units of text involved in any NLP task. This is the first step when dealing with textual data. By splitting into tokens it becomes easier to focus on only the relevant data. This can be done using word\_tokenize. We can also detect words containing only alphanumeric characters using regexp.

Removing stop words: very commonly used words such as the, is , he are called as stop words. Not only English but other languages also have similar words. Stop word removal removes words that do not hold any relevance to us. It can be done using stopwords.stop(‘english’).

Stemming: This is a text preprocessing task where we convert or transform related or similar words to their base values because they share the same meaning. This can be done using stemmer.stem() function. However stem shortens the words into random characters of no specific length or words that may not even make sense to us. Like for example computer can become comp, comput or anything random.

Lemmatization: Lemmatization is stemming but more powerful. It not only reduces the words but considers a languages full vocabulary and then returns the root of the word. Hence unlike stemming it is better to use lemmatization for standardization.

Handling Contractions: Expanding contractions such as cant to cannot.

Normalization: To ensure consistency in the textual data normalization is done such as converting abbreviations to full words.

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.