Task 3.1

1. First we need to **remove all the null values** from our dataset i.e. messages, emails, books, etc. This can be done using **dropna()** function which drops the rows having null values.
2. **Remove punctuations and special characters** such as: (,), “,”(comma), “.”(full stop), #, @, etc. This reduces noise in the text
3. **Convert all the text into lowercase** because it makes the text uniform and gives a consistent output. And two different words with different capitalization will be treated similarly or as same tokens. Eg: “school” and “SCHOOL”
4. The we will **remove the duplicate words** so that computer is able to identify unique words
5. **Tokenization** => breaking the text into different tokens, where each token can be a sentence, a word, or just a single character. It helps computer to break down human language into smaller units which are easy to understand.
6. **Removing stop words** => Stop words are words that are commonly used in a language and are not required to understand the context of the text, they have no or very little meaning such as: ‘a’, ‘an’, ‘the’, ‘to’, ‘so’, ‘or’, etc… This allows NLP model to focus on important words which has meaning and adds context.
7. We will combine all the relevant words, that are not stop words in a corpus list.
8. We will perform a process of text transformation. There are two types of transformation which I know:
    a. First, **Stemming:** We reduces word to root form even if its actually losses its meaning. Eg: running => run, caring => car, achieve => achiev and achieving => achiev. It is less accurate but faster.
    b. Second **Lemmatizaton:** We reduces the word to base dictionary form where the actual meaning of the word is maintained. Eg: caring => car, achieve => achieve and achieving => achieve. It is more accurate but slower.

Task 3.2

Clustering

1. It comes under **unsupervised learning**.
2. It refers to making **groups of unlabelled data sets** known as **clusters**.
3. We **group similar data** to make sense of it and use it for various purposes like targeted ads, recommendations, etc.
4. These data groups are based on similarities among the unlabelled data sets.
5. **Streaming platforms** recommend movies based on watch history, clustering groups of users with similar viewing patterns.
6. **Banks** use clustering to group customers into different risk categories based on their financial history, which helps decide loan eligibility.

There are many types of clustering. Two of these are:

1. **Partitioning Clustering**:
    a. Here we initialize centroids in the data set and then each data point is compared with the centroids initialize.
    b. The centroid closest to a particular data point is associated with that data point and all the data points associated with a centroid forms a cluster.
    c. And the mean of the data point is taken inside a cluster to create new centroids and then the above process is repeated until we get unchanging centroids.
    d. It is also known as K-Means clustering, where K refers to the number of centroids initialized.
2. **Hierarchical Clustering:**
    a. It builds a tree like structure of nested clusters also known as dendrogram. It is of 2 types:
        i. **Agglomerative (Bottom-Up):**
            1. At start each datapoint is a cluster and then every cluster is merge to form a bigger cluster until a stopping criterion arrived or only one cluster remains.
        ii. **Divisive (Top-Down):**
            1. Start with all data point in one cluster and then it starts splitting until a stopping condition is met or a single data point is a cluster itself.