Task 3.1

Let’s say you are given a large amount of textual data- messages, emails, books, etc. Before performing any operations on this data, it is necessary to clean and preprocess the data (removing unnecessary words or symbols, etc.). Explain how you would go about preprocessing. What different steps would be followed? Why are they necessary?

Data preprocessing should mainly include following steps, but before we get into preprocessing lets also understand data collection:

* Data Collection

The data you might need in order to train your ML model or even just for data analysis comes from here. This is the initial mark for further steps, which makes this step crucial.

* + The data to be collected should come from verified and trustworthy sources.
  + It can be found in databases, SQLs or even in CSV, JSON and XML file formats.

Data preprocessing is a foundational step when it comes to converting raw data into some sort of useful information. The raw data can be noisy, faulty, incomplete or missing some values.

Steps involved in data preprocessing:

* + Data Cleaning
  + Data Transformation
* Data Cleaning and Data Integration

The data you would work on will not always be clean and workable with each and every detail up-to the par for your project, so before you start working on it, you must clean the data. Cleaning refers to following tasks:

* + Taking care of missing/empty values.
  + Handling incorrect data types and converting them into appropriate ones.
  + Removal/Replacement of invalid values.
  + Eliminating irrelevant data.
  + Elimination of duplicate data.

The process of preprocessing directly translates into the reliability of your model. These steps ensure that our ML Model or Data analysis will work with relevant, filtered and comparatively better data quality and thus enhancing the quality and performance of our model.

Data is not always collected from a single source, it might be collected from multiple sources, thus it is important to integrate filtered data from multiple datasets into a single unified dataset.

* Data Transformation and Data Reduction

Data collected can come in variety of formats, so it is necessary to transform all data into a suitable format for data analysis and ML modelling. The end goal of data transformation is to make data more accessible, understandable and use it up-to its full potential. Some key elements of data transformation are:

* + Transformation techniques: Transformation techniques are methods or operations applied to the data to alter/modify data’s format, structure and representation. Some major techniques are Normalization, Standardization, One-hot encoding etc.
  + Feature engineering: Even after filtering one might not use entire dataset, but rather only relevant data. Feature engineering mainly consists of creating new features from existing data or only considering relevant data for analysis and modelling to improve model performance.
* Data splitting

Data splitting is when the given data is divided into two or more subsets so that a model can get trained, tested and evaluated. If we have 2 splits, one will be used for training while the other would be used for testing, and if we have 3 splits, there will be training, testing and validation sets.