Question - 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?

Solution-

Preprocessing in general, includes several tasks, such as cleaning data, handling missing values, normalizing or scaling features, and reducing dimensionality. Each step helps refine the dataset so that the machine learning algorithms can interpret the data correctly and efficiently.

Data Preprocessing can be done by using panda library of python, or scikit learn.

There several ways in which textual can and need to be pre-processed, some of them are-

1. Lowercasing – such that different instances of the same word can be treated as same. This will also help in reducing the no. of conditions to put. Eg. Text, TEXT, text will all be treated same when converted to lowercase
2. Removing Punctuation - This is again a text standardization process like lowercasing that will help to treat 'hurray' and 'hurray!' in the same way.
3. Removing Stopwords (they are unnecessary – for,etc)
4. Removing emojis/emoticons (should not be done for sentiment analysis tho)
5. Spelling Correction – so that we don’t get any meaningless words in our database
6. Removing rare/frequent words that won’t be of much use to us in certain cases of data processing
7. Either Stemming or Lemmatization –
8. Stemming – It helps in reducing a certain down to its root or base word. However it can not always be write for e.g. if the word is talking it can reduce it to talk but for consoling, it might get reduced to consol which is not an actual word.

The most famous way for stemming is to use PorterStemmer in nltk.stem.porter.

1. Lemmatization – this is similar to stemming but it actually makes sure whether the reduced root word actually belongs to the language. However, it is slower that stemming process. So, if there is a speed we have to choose between lemmatizing or stemming.

Lemmatizing can be done by using WordNetLemmatizer in nltk

1. **Important for Chat Analysis –** Chat Word Conversion – while chatting people often use abbreviated forms of words. These words have to be converted to their original form for smoother analysis

**What is the necessity?**

Surely we can use raw data to analyse something, but its output might just turn out to be bogus answers that will further complicate the entire understanding process and will reduce the entire algorithms efficiency.

Pre-processing the data is important so as to avoid the bogus or wrong answer situation and to get efficient, compact and understandable outputs. In short, preprocessing enhances the data's quality and makes it easier for machine learning algorithms to read, use, and interpret it.