Synapse Machine Learning 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?

For preprocessing a large amount of textual data like messages, emails and books etc. for it to be ready for analysis the following steps should be taken:

* Converting all the text into lowercase – Through this step we convert all the data we have into lowercase using the .lower() method. This makes analysis easier as it makes “Text” and “text” into all lowercase and helps maintain consistency.
* Removing punctuation marks – Punctuation marks usually hold no meaning so it is better to remove them for smoother analysis and ease further steps. It is done through the re.sub() method.
* Tokenization – After removing the punctuation marks we are left with a paragraph of words which we have to break down into smaller parts to facilitate removal of stop words or stemming/lemmatization. Helps understand the text at word level and can also be taken to the character level.
* Removing the stop words – Next comes the removal of stop words like the, and, is etc. from the given text. This helps in removing the frequent stop words which do not add much value to text analysis. Their removal helps in focusing on other much more important words.
* Stemming or Lemmatisation – Once we have broken down the given text into smaller parts like phrases, words or characters we start a process called stemming or lemmatization in which we try to reduce the dimensionality of the text by removing the root word or the common part from similar words for example we get ‘run’ from running, ran, runs etc.
* Expanding Contractions – In this step we separate the contracted words such as isn’t or don’t into their root words, is not and do not. This helps us maintain consistency in the text.