Task 3.1

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?

Text Preprocessing — Text preprocessing is the first thing you should do in processing a task using Natural Language Processing methods (NLP). It involves cleaning and converting unstructured text data into structured format for analysis. This is how more sophisticated NLP models that come after are trained and perform better.

Text Preprocessing Steps:Lowercasing

In order to keep everything consistent and not have the same word inputs treated differently just because of capitalization differences your text should be turned all lowercase.

Eliminating Punctuation:

Removing punctuation will generally make text analysis simpler, and create fewer noisy features But selective removal maybe necessary though depending on the work some of those punctuation could be essential.

Elimination of Stop Words:

Stop Words: words such as "the", "and" and (some would argue) even the word “is” seem to have no meaningful value. This can decrease the model dimensionality and improve efficiency by removing the meaningless (semantically speaking). By removing them we can make the model a lower dimensionality and more efficient.

Tokenization:

In many NLP tasks, breaking text into words (language modeling) or tokens is critical. It allows for more to be processed and analyzed.

Lemmatization vs Stemming :

When words are down to their root form, it is more easy to identify them and text representation can also be improved.

Taking suffixes off the ends of words (e.g., turning "running" into run →)

Linguistic Processing: Lemmatization- it is the process in which you convert a word into its base form based on morphological analysis and provides with vocabulary(eg—better becomes good)

Depending on the assignment, numbers may or may not be that significant. Select whether or not to convert numbers into text._RESULTS.shuffle().

Strip special characters:

Remove special characters such as emojis, symbols and control characters which are not useful for the analysis.

Managing Infections:

While the extension of contractions would help in better representation of text, when done mindlessly it can introduce noise.

Importance

Why text preprocessing is needed, well it helps in improving the performance of models. Processing the data further will bring an additional increase in accuracy and efficiency of your model.

Reduction in the noise: By removing the paradoxical and irrelevant information you improve on quality of data.

Uniform representation: Normalization and lowercasing returns a standard (normalized) format for analysis.

Feature Extraction: To convert preprocessed text into numerical representations for ML algorithms.

More Compute: Smaller data set, and/or easier features to process makes for faster computation.