Natel Whitaker
SER 401
Individual Research: Data Preprocessing and Data Annotation

From my research into the topic of data preprocessing and annotation, I was able to ascertain that at its core; they aim to change the raw data into something more suitable to use on an LLM. For the data preprocessing aspect, the main goal is to ensure the data is consistent and high quality throughout. This could be done through tokenization, assuring all textual data is of the same case, or handling data with special characters. Annotation, on the other hand, is then labeling that data with a sort of tag or information that the model can then use to learn from.

Useful links: Preprocessing Steps for Natural Language Processing (NLP): A Beginner's Guide | by Maleesha De Silva | Medium

Text Annotation for NLP: A Comprehensive Guide [2024 Update] (habiledata.com)