Notes on NLP

* Language Model is a model which can predict the next word in a sentence. It can be trained on a corpus of data like Wikipedia.
* How do we build one ?
  + Tokenization
  + Numericalization
  + Language Model Dataloader Creation
  + Language Model Creation
* What are the independent variables and labels in a Language Model?
  + As mentioned in Point 1, we are trying to predict the next word.
  + Using this information, independent variable will be the sequence of words starting with the first word in the long list of words and ending with the second to last, and our dependent variable will be the sequence of words starting with the second word and ending with the last word
* Tokenization is splitting sentences into words using spaces as delimiters (general approach). It can also split based on sub-words or even individual characters
* Numericalization is mapping the tokens to numbers.
* As mentioned earlier, the X and Y for Language Models are something like mentioned below
  + X = What is this issue that we are Y = is the issue that we are seeing
  + A dataloader has to be built with X and Y like this.
* Language Models use a Recurrent Neural Network. Perplexity is a metric used and cross entropy is the loss function used generally.
* This Language Model can be used to create a Text Classifier (sentiment analysis)
* However, for this we need labels. Language Model did not need one since we predict the next word which is also present in the corpus.
* Using the labels and the text for the classification task and using the vocab from Language model and the encoder from Language Model, we can train a text classifier.