For data in general, a majority of it is almost unstructured data at this moment. It can come from multiple sources such as natural language processing or simply pulled right off the Internet. But once you have obtained the source of the data, you have some options in how to prepare it for analysis or apply it to a machine learning algorithm. The first option is to perform text cleaning. In this step, you should review the text and look for any anomalies that may be present within it. (Brownlee, 2019) An example would be how punctuation is being used or if there is any extra characters used, or an interesting case would be if there was a translation of the document into UK English where some words will be slightly different than what you are used to seeing your text. The next option someone could use for cleaning the data would be to enable manual tokenization. (Brownlee, 2019) In this case, you would load in the data then split the data on whitespace followed by punctuation. Looking at another option, you could normalize all the text from the source. Like the tokenization, you will load and separate by whitespace but instead of removing punctuation you will make all of the text lowercase which would be to normalize the entire data as one case. Author Jason Brownlee brings up how this could be an issue with names of companies or products that may have the same name as an object such as Windows and windows. The options that I have mentioned above are simply a few paths someone could take in cleaning the data. With these options, you also have to understand how to handle extracting text from multiple sources as well as working with numbers that could relate to either currency or the calendar. When it comes to data cleaning with unstructured data, it sometimes will never be a hundred percent clean but the goal it to clean it up as much as possible so that it does play a factor down the lines when applying it to a model.

In simple, the bag of words model is a way of representing text data when modeling text with machine learning algorithms. (Brownlee, 2019) With this model, it usually consists of two major factors. The first being a vocabulary of known words and the second is a measure of the presence of know words. With the bag of words model, its primary concern is with the number of occurrences of words; it has no association with how the sentence or document is actually structured at the time.

However with this model like in all models, there are some limitations to the bag of words. The main three areas are vocabulary, sparsity, and meaning. (Brownlee, 2019) With discarding how the sentence is structured within the document, I think it loses its potential to further assist the model. For a lot of words, they practically have the same meaning simply a different way of expressing it.

References:

Brownlee, J. (2019, August 7). A Gentle Introduction to the Bag-of-Words Model. Retrieved March 18, 2020, from <https://machinelearningmastery.com/gentle-introduction-bag-words-model/>

Brownlee, J. (2019, August 7). How to Clean Text for Machine Learning with Python. Retrieved March 18, 2020, from <https://machinelearningmastery.com/clean-text-machine-learning-python/>