**Data Cleaning**

During various stages, data was cleaned or prepared for analysis. During ingestion, the Federal Reserve policy action had some missing values that needed to be input. For example, when there was a meeting where there was no action, the “Level” variable would have an empty value, which would need replacing with the last observation to keep the Level steady. If there was a meeting that had a decrease in rates, this would imply no increase in rates. Thus, 0s were input for those missing values. In unemployment data, some missing values were replaced with the last observation as well, as in time series analysis, it would not be right to replaced these with the mean, as a skewed graph would lend itself to a biasedly higher or lower mean, and would not be representative of the missing value. For data preparation, the text also had to be cleaned in preparation for vectorization and tokenization. The process first includes reducing all words to lower case. This is important for “normalizing” words, in a sense, so that they are all valued the same. Punctuation is then removed to identify the words separately and without need for context. Stop words are filtered out, which decreases the noise from words with no specific value such as “the”, “is,”, and “and”. The list of word are then put through a “lemmatizer”, which reduces the words to their base form (for example, from “running” to “run”). This has the effect of standardizing many words in different forms that essentially mean the same thing. Then, all of the text is joined again by row according to the Minutes they came from. This forms the variable of “Preprocessed Text” which will act as the data through which vectorization and frequency analysis will be generated.