1. Feature set is in processed.txt file and is generated by the Sentence\_Processor.py file.
2. The Term Document Matrix is generated by the Sentence\_Processor.py file.
   1. Tokenizing the sentences is needed in order to process each individual word. No information is lost in the process.
   2. Removing punctuation allows for words to be identified and matched. It could potentially cause some words that aren’t the same to be mistakenly classified as duplicates.
   3. Removing numbers does not alter the “meaning” of a sentence. However it does effect the magnitude of some words or phrases.
   4. Converting upper-case letters to lower-case allows for words to be matched if they have a capital letter in them. This has no real effect on the data.
   5. Stop words are removed because they have no effect on the meaning of a sentence.
   6. For some reason the stemming algorithm provided changed some words that it shouldn’t have. For example every word that ended with an “e” had is trailing “e” removed. A similar case happened to some singular words like “anonymous” and their trialing “s”.
   7. Stemmed words were combined in the process of forming the TDM, they do not cause the loss of any information because they are effectively repeating the same word. However repeats are still accounted for in the TDM to make sure that no data is lost.
3. In the TDM you can see that there are 395 columns of data, the same is true of the reduced TDM where there are 33 columns of data. As a result of this there was no clear way to apply a clustering algorithm that would have been effective. Perhaps finding some way to classify similar words as the same thing would reduce the number of columns but it didn’t seem like it would make a drastic difference. As a result there was no effective way to classify the data in the TDM.