Report on PA1

Tokenize:

Initially when we tried to implement the tokenizer, we tried using the Stanford CoreNLP. The problem with the core NLP is that returned information that was too in depth and included information on character offset and part of speech, which needed to be parsed out. After consulting with Tuan, our group decided that Core NLP would take too long to run and that we would write our own tokenizer. The tokenizing stage was relatively easy and we used the tokenizer class after taking out stop words and punctuation and different markups. After writing our own lemmatizer, we tested it on a book, *Made to Stick*. We tested it around 5-10 times and adjusted the lemmatizer every time until we thought the lemmatizer worked pretty well.