Summary:

I created a text generation system that creates the title and text of a Wikibook. I used a LSTM neural network for my model and used a variety of different nltk and keras tools for cleaning and preprocessing my data. For cleaning, I used regular expression, stopwords, spacy, and string manipulation to filter out unnecessary data from my text. To prepare my data for the neural network I tokenized my text using keras, converted that text into a padded sequence, and then fed that into my LSTM. My LSTM comprised of only 3 layers, 1 embedded layer as my input, 1 LSTM layer for my hidden layer, and a softmax dense layer as my output.

Challenges:

A big challenge for this project was having enough gigabytes of memory to create a padded_sequence with a large dataset. I had to cut the dataset used for the body by quite a lot as my PC didn't have enough memory to perform it. Another difficult part of this project was having the generated body text make any amount of sense. I tried to fit the model with and without stopwords, and with different cleaning methods to get different results, but it all ended up being nonsensical. For the scope of this project, I couldn't figure out how to keep the model on topic.

Conclusion:

I'm happy with the results, but I think using this method of text generation for the text body was a mistake. It's too imprecise and doesn't understand sentence structure and patterns with language. I think having AI creating a whole paragraph of text is beyond the scope of this project, as it needs more data for better word selection, a model to understand sentence structure, and to be able to stay somewhat on topic after the first few words. Nonetheless, it was a fun project and I learned a lot about how to use text from a dataset in a LSTM neural network.