Names: Sean Lee & Maria Mercado

Brief Description: Compare the effectiveness of different encoder-decoder models which can take wikipedia articles and their corresponding titles and reproduce those articles given their topic. The 3 models that will be compared are:

- A transformer model built from scratch
- Pretrained Bart transferred onto our dataset
- Non-attention based model (LSTM)

The goal of this comparison is to judge the effectiveness of attention in generating long outputs from a short input. These methods can be applied for summarization and compression(since we are trying to create a long sequence from a short one.

Resources:

On the general problems with generating wikipedia articles:
https://towardsdatascience.com/generating-wikipedia-articles-with-ai-995436c9f95f
Transfer learning using Bert implementation:
https://qa.fastforwardlabs.com/pytorch/hugging%20face/wikipedia/bert/transformers/2020/05/19/Getting_Started_with_QA.html
Bart Paper:
https://arxiv.org/abs/1910.13461
Attention Paper: https://arxiv.org/abs/1706.03762

Project Plan:

1. Create a dataset from scraping wikipedia pages. Benefit is we can easily do this and control the size of our dataset. It will be more useful to have many pages looking only at the first few paragraphs than a few pages but long articles.
2. Create Custom embeddings. Decide on a tokenizer(either Glove or Bert) and retrain it on our dataset.
3. Create non-attention(LSTM) and pre-trained models for comparison.
4. If possible, transformer model will be trained from scratch on SCC.
5. Evaluation metric will be how well the model is able to recreate the original article given the title.