CS685 Quiz 4: ***text generation***

Released 10/5, due 10/8 on Gradescope (please upload a PDF!)

Please answer both questions in 3-4 sentences each.

1. **Give three reasons why we might prefer a character-based tokenization over a word-based tokenization when training a neural language model.**
   Character based tokenization may be preferred over word-based tokenization when training a neural language model because -
   a. It reduces the size of vocabulary
   b. There is a very low chance that a character token may be out-of-vocabulary
   c. Misspelled words can be corrected rather than marked as unknown in a character tokenized model.

2. **Assume you are given a sequence-to-sequence model trained to solve the task of *paraphrase generation*, in which a model receives a single sentence as input and is asked to produce a paraphrase of that sentence (i.e., a sentence with approximately the same meaning). Why might you want to decode from this model using nucleus sampling instead of beam search?**
   Nucleus sampling would be preferred over beam search for paraphrase generation because chances are that the model using beam search may produce the same sentence as input as it searches for and returns the most likely text. To generate a paraphrase with similar meaning, we would need more diversity in the text generated from our model. The nucleus sampling method generates text by sampling the dynamic nucleus of the probability distribution, hence allowing for more diversity in the generated text but still retaining the likeliness and coherence.