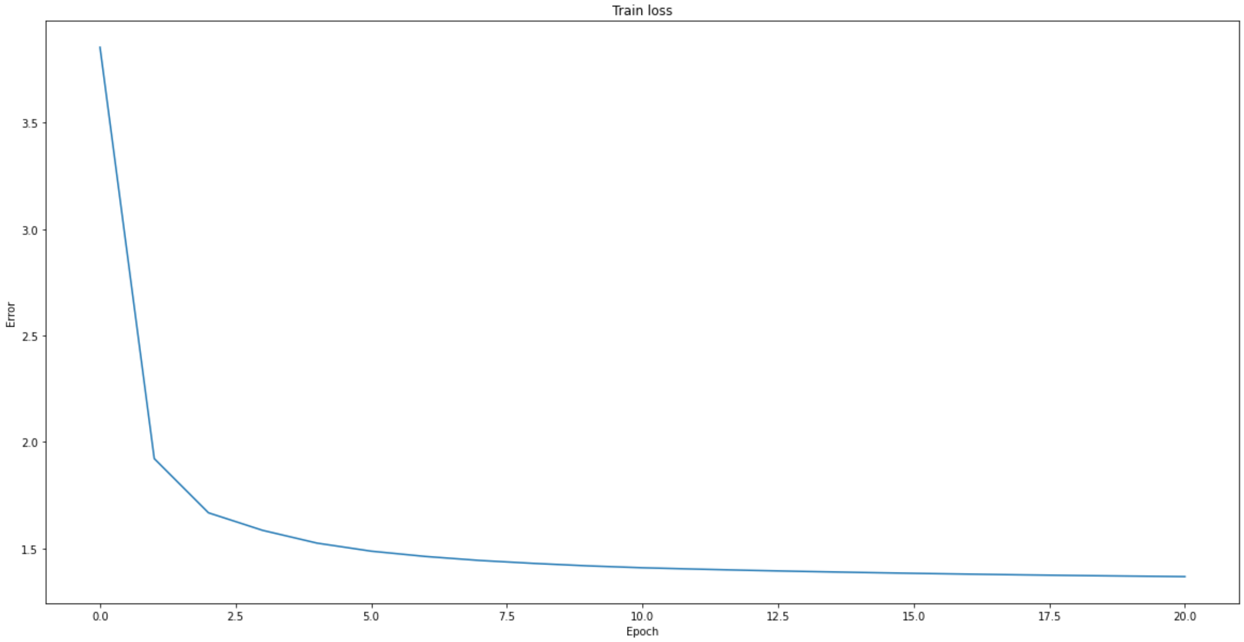
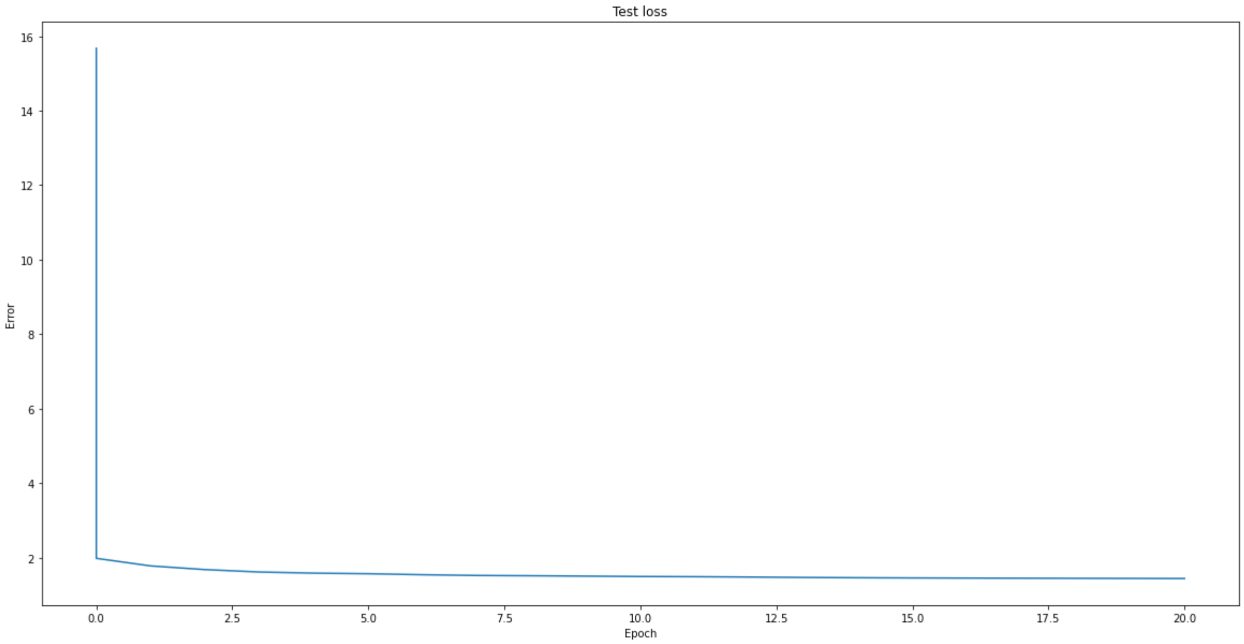
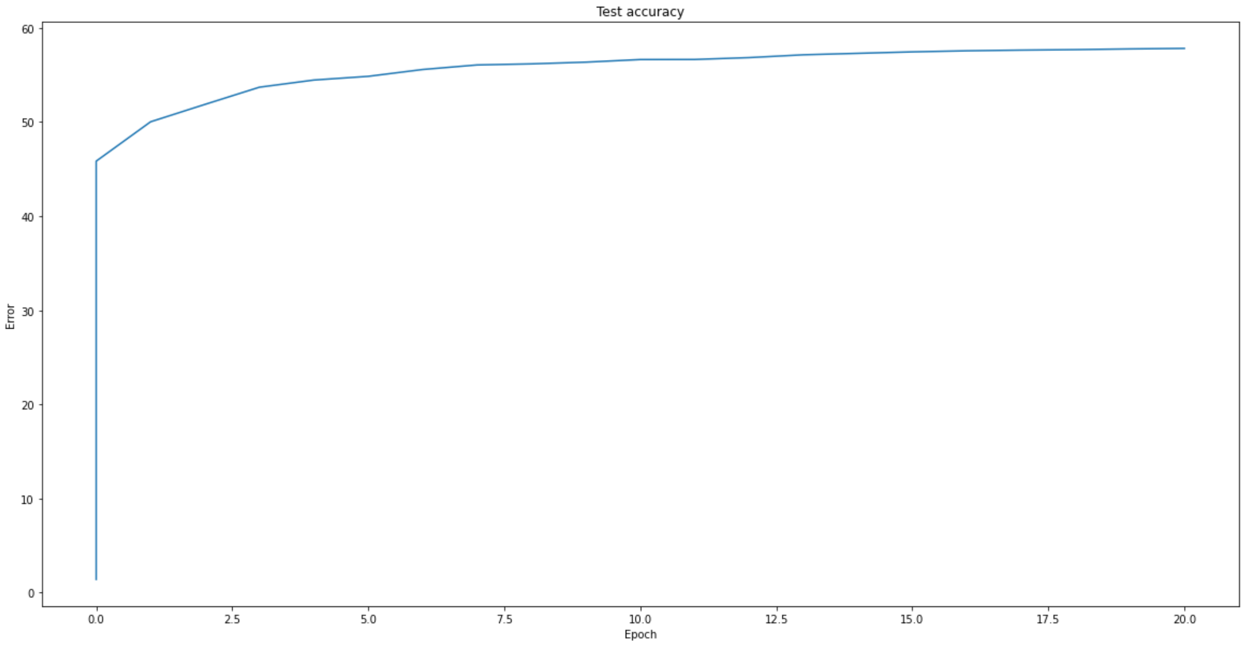
# Part 9: Short answer questions

Please answer these questions, and put the answers in a file called short_answer.pdf in your repository.

1. Just like last time, provide plots for training error, test error, and test accuracy. Also provide a plot of your train and test perplexity per epoch.
    - In class we defined perplexity as 2^(p*log_2(q)), However the PyTorch cross entropy function uses the natural log. To compute perplexity directly from the cross entropy, you should use e^p*ln(q).
    - We encourage you to try multiple network modifications and hyperparameters, but you only need to provide plots for your best model. Please list the modifications and hyperparameters.







2. What was your final test accuracy? What was your final test perplexity?

The final test accuracy is 58%. Since the final cross entropy loss is 1.4358, the final perplexity is given by e^1.4358=4.203

3. What was your favorite sentence generated via each of the sampling methods? What was the prompt you gave to generate that sentence?

Max: Harry Potter and the corridor was staring at the corridor and the corridor was staring at the corridor and the corridor

Sample: The love between the wand and shake leftrame. "t Gryf?" said Harry though candled on the book bac% out of their arg

Beam: The love between that they were staring at Harry and Hermione and Hermione was staring at Harry.

The underline is the prompt given to each method.

4. Which sampling method seemed to generate the best results? Why do you think that is?

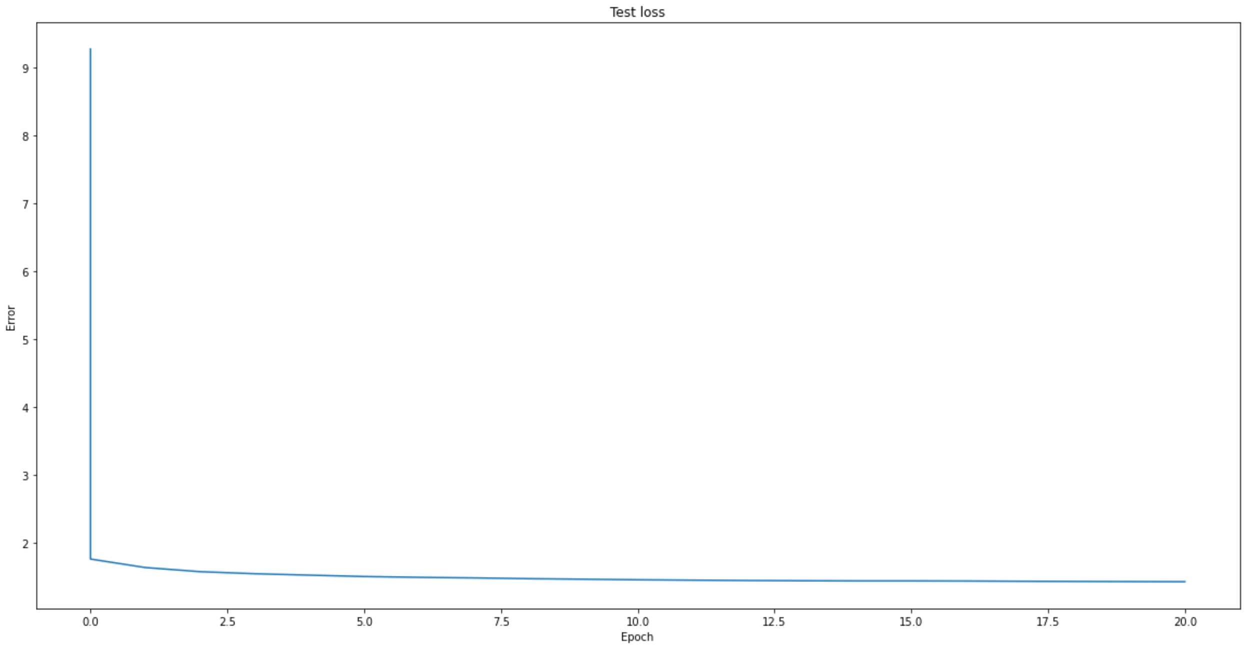
   In a way, Sample seems to generate the best results. This is because the other two can easily produce loops.

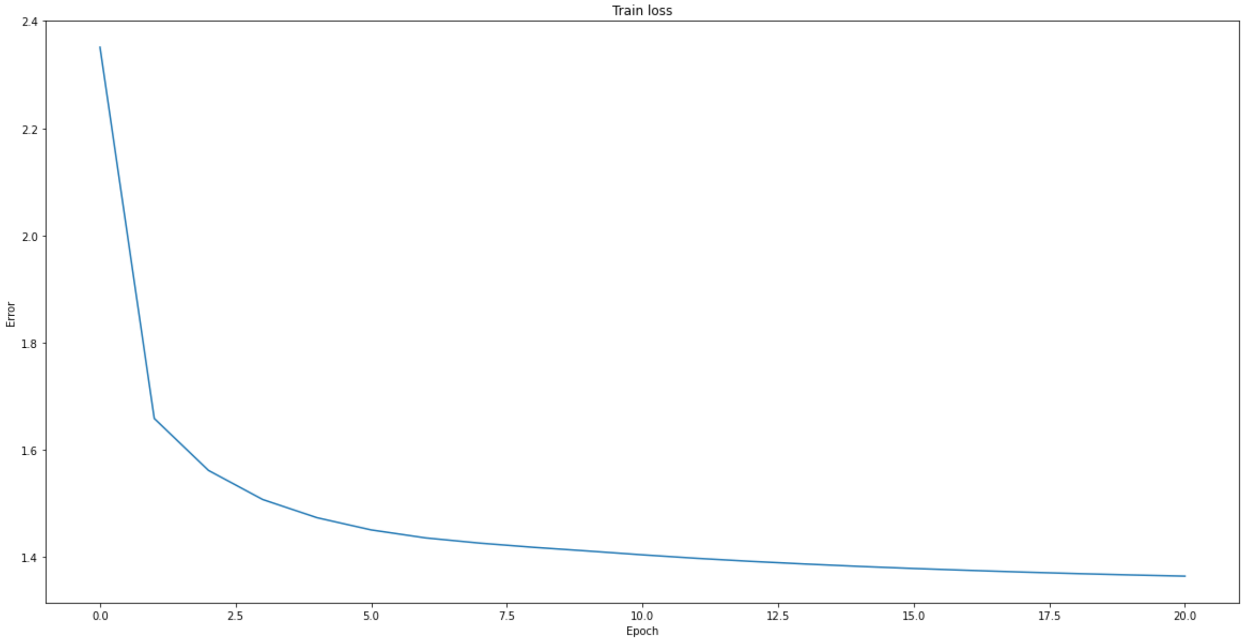
5. For sampling and beam search, try multiple temperatures between 0 and 2.
   - Which produces the best outputs? Best as in made the most sense, your favorite, or funniest, doesn't really matter how you decide.
     i. Temperature=0.5 makes a really grammatically correct and understandable sentence, meaningless though.
   - What does a temperature of 0 do? What does a temperature of 0<temp<1 do? What does a temperature of 1 do? What does a temperature of above 1 do? What would a negative temperature do (assuming the code allowed for negative temperature)?
     i. A temperature of 0 would cause all the output to become infinity, potentially making every entry equally likely.
     ii. A temperature of 0<temp<1 would cause the sampling to be less random; values that are higher are going to have a much higher probability than those that are lower.
     iii. A temperature of 1 would just return the softmax distribution.
     iv. A temperature greater than 1 would cause the probability distribution to be flatter; each output is going to be more equally likely.
     v. A negative temperature would inverse the probability distribution; making the largest output to be the least likely.

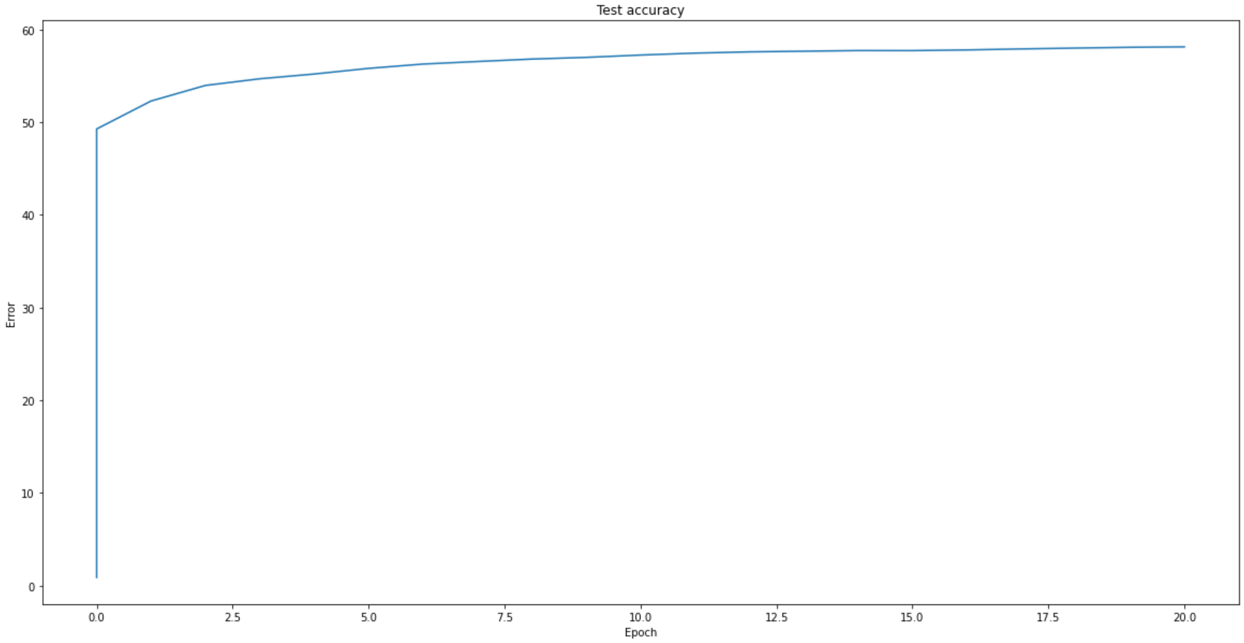
Questions for each of the "Other things" sections. Only answer the questions corresponding to the ones you chose.

1. New Corpus
    1. What corpus did you choose? How many characters were in it?
        i. We chose the bible. It has 4251004 characters.
    2. What differences did you notice between the sentences generated with the new/vs old corpus.
        i. The new corpus sentences always have numbers in them as is for the bible. The new corpus uses a lot of religious words while the old uses a lot of fantasy words.
    3. Provide outputs for each sampling method on the new corpus (you can pick one temperature, but say what it was).
        i. Temperature is 0.5
        ii. Max: Love is the tenth one of the seen of the LORD of hosts. 1:11 And they shall be the desolate the temple that
        iii. Sample: Love is the land. 51:33 And Egypt me came unto full my destrylets, and mell and he said unto them, 19:5J! B
        iv. Beam: Love is into the house of the LORD of hosts; and they shall come to pass, that they shall come to pass

2. New Architecture
3. LSTM
    1. What new difficulties did you run into while training?
    2. Were results better than the GRU? Provide training and testing plots.







It’s not significantly better.

    3. Provide outputs for each sampling method on the new corpus (you can pick one temperature, but say what it was).
        i. Temperature: 0.5
        ii. Max: Harry Potter, Voldemort, and Dumbledore walk into a bar. Harry stared at the corridor and staring at the corridor and staring at the corridor and staring at the corridor and staring at the corridor and staring at the corridor and staring at the corridor and
        iii. Sample: Harry Potter, Voldemort, and Dumbledore walk into a bar. looked at Harry's golden larvizade polder insceasancl of ragenry about the came as she had not aunt few withten nose, trying to quite the carest behind him his tands, so;rounded the dark and he taiked
        iv. Beam: Harry Potter, Voldemort, and Dumbledore walk into a bar. Harry was staring at Harry and Harry and Harry and Harry and Hermione was staring at Harry and Harry and Harry and Harry and Hermione was staring at Harry and Harry and Harry and Harry and Hermione wa

4. Transformer
5. Student-forcing
6. Words
    1. What new difficulties did you run into while training?
        i. Word creates a much larger vocabulary.
        ii. Punctuations have to be stripped to create better vocabulary.
        iii. Words have to be turned into lower cases.
    2. How large was your vocabulary?
        i. I tried to include every word that appeared more than 5 times, which resulted in memory overflow in CUDA.
        ii. Then I tried 30000 words which also resulted in memory overflow.
        iii. The running code used 10000 words.
    3. Did you find that different batch size, sequence length, and feature size and other hyperparameters were needed? If so, what worked best for you?
        i. I find shorter sequence lengths have better performance. I reduced the sequence length to 32.