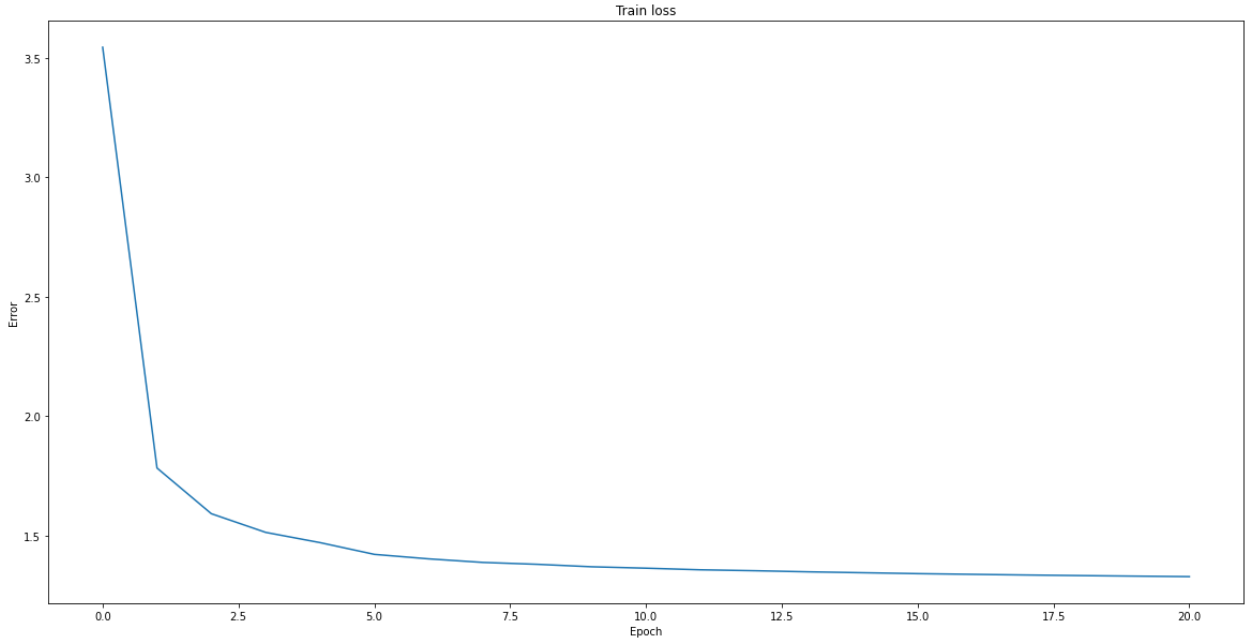
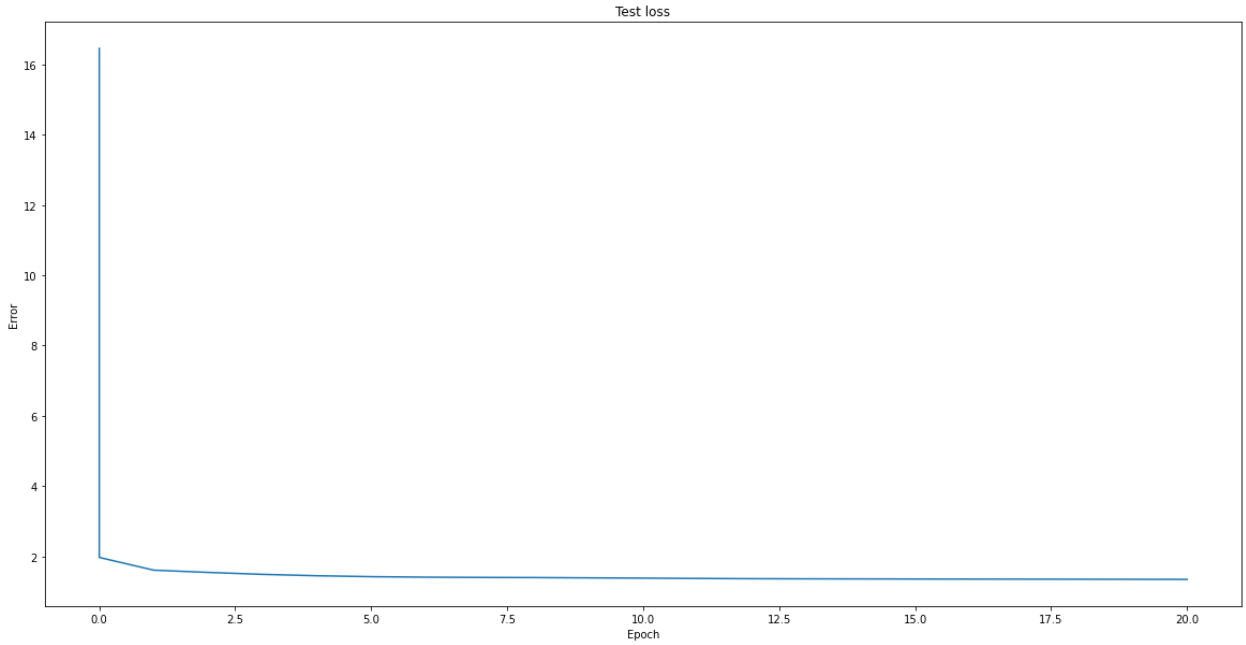
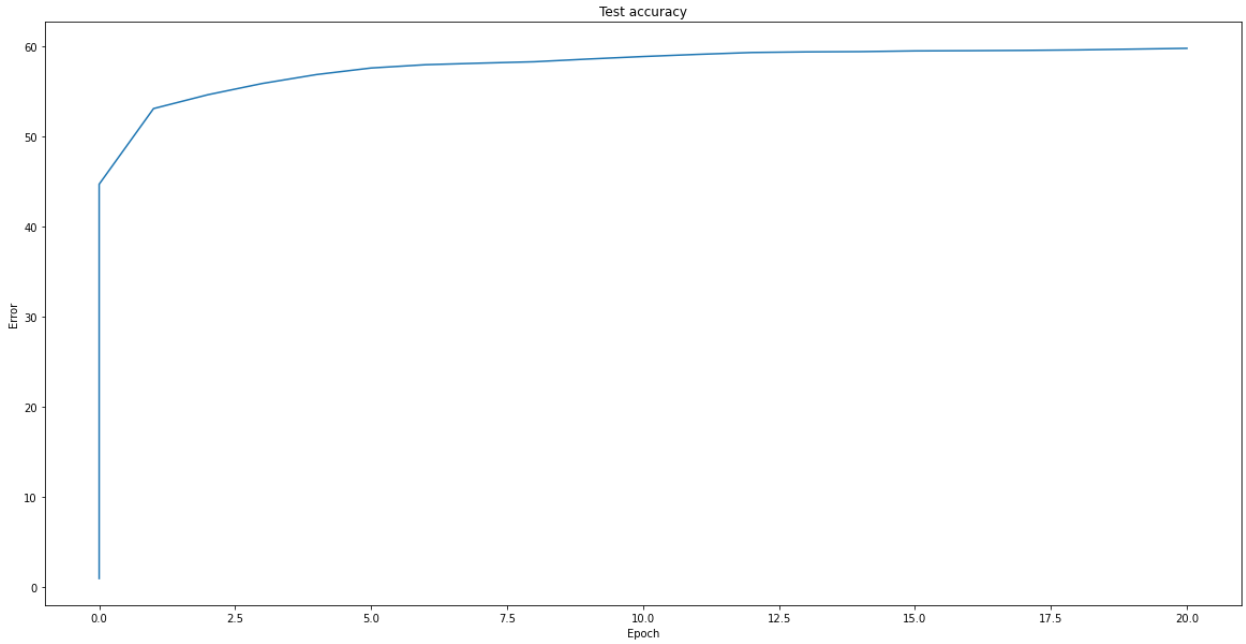
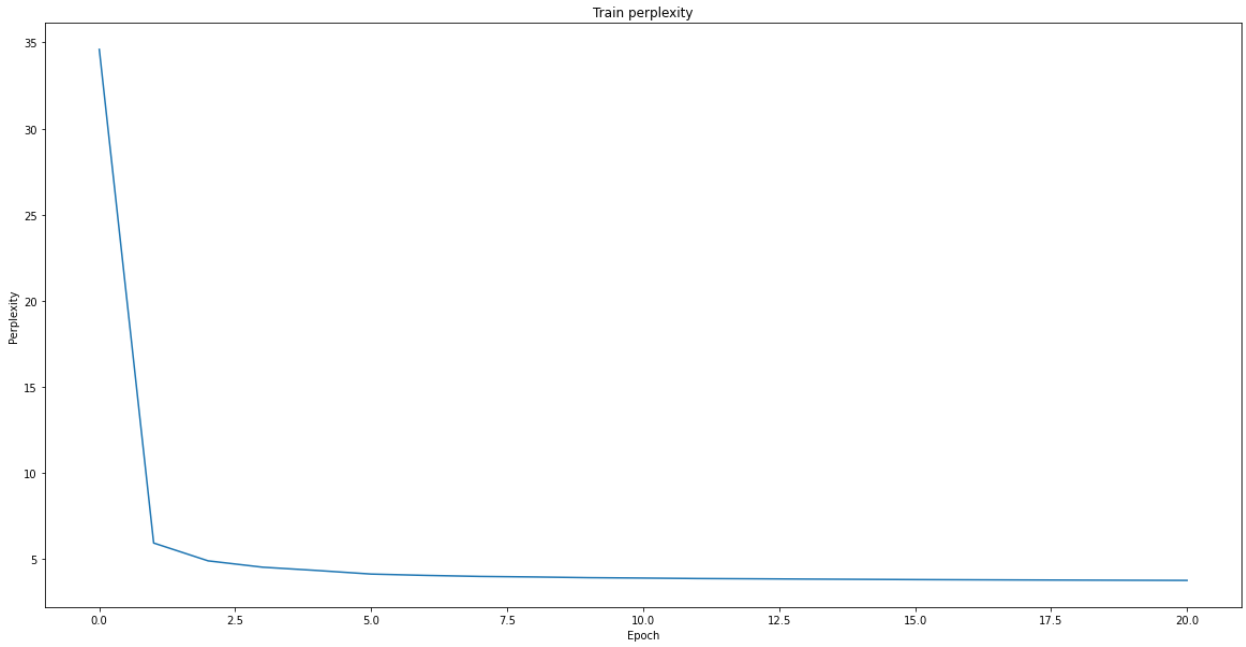
Deep Learning HW2 Short Answers

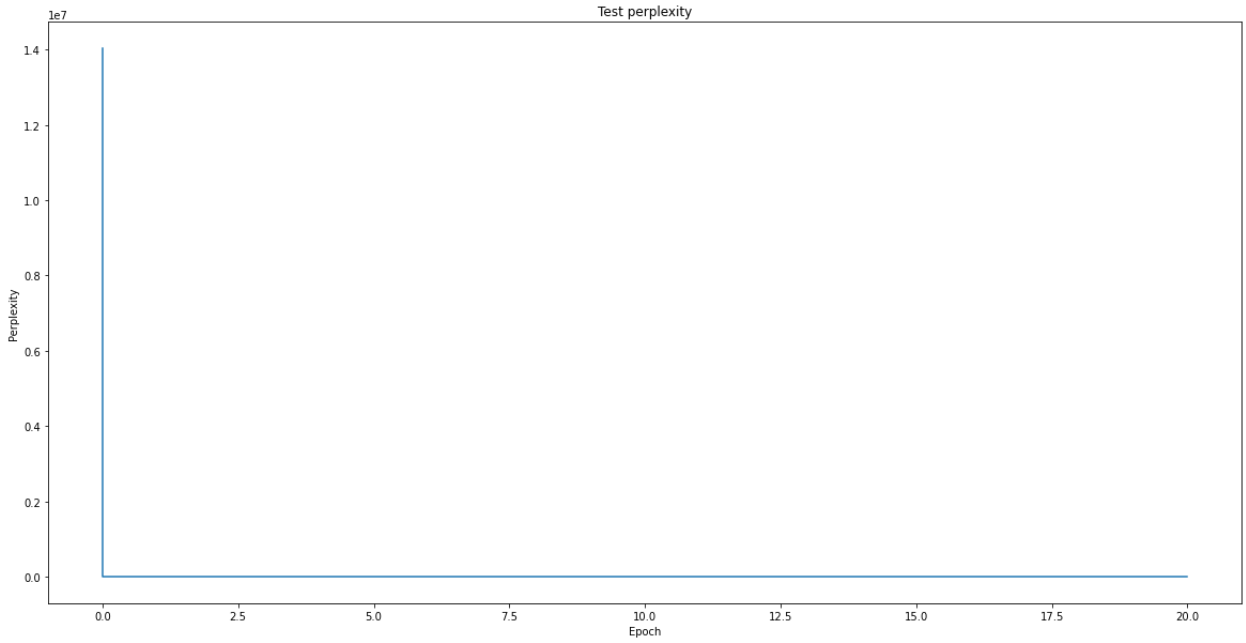
1. **Just like last time, provide plots for training error, test error, and test accuracy. Also provide a plot of your train and test perplexity per epoch.**






Test accuracy
Error
Epoch
0
10
20
30
40
50
60
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Train perplexity
Perplexity
Epoch
5
10
15
20
25
30
35
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0



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

Final Test Accuracy: 59.75%
Final Test Perplexity: 3.82

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

Max Sampling: Harry Potter and the bottom of the floor and said, "I don't know what was all the point of the parchment when he was all right and started to the back of the corner and said, "I said the portrait to the parchment when he
Prompt: “Harry Potter and the”

Sample Sampling: Harry Potter and there was a small snappy of the pain flamed and looked at the wall. He looked up at Harry and Gonder and Hermione had already was all saying and they was all stupped on the dark and was still thinking an
Prompt: “Harry Potter and the”

Beam Sampling: Harry Potter and the bottom of the floor and said, "I don't know what was all the point of the parchment when he was all right and started to the back of the corner and said, "I said the portrait to the parchment when he
Prompt: “Harry Potter and the”

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

The sample sampling method seems to generate the best results in terms of the variation in what is produced. The generated text from max sampling and beam sampling start off without making too much sense, and the texts, although may make sense in a window of 3-4 words, don't make sense in the context of the entire generation. On the other hand, the sample sampling, although it has some typos, seems to put together a sentence with more substance behind it. This could be due to the fact that the sample sampling method is heavily influenced by temperature during inference, due to the fact that we sample from the model output distribution and hence with a low temperature such as 0.75, we are able to generate a good balance of elements that make use of the model's understanding of the corpus and random choice.

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.

A temperature of 0.75 seems to produce the best outputs in the context that they make sense and have at least localized meaning associated with them.

- What does a temperature of 0 do?
    - A temperature of 0 will result in more confident predictions because the larger unnormalized probabilities will have even higher softmax values, and therefore, the normalized probabilities will be much higher than when using nonzero, positive temperature values. Since we are actually dividing by temperature, this should actually result in an undefined output, but as temperature gets closer and closer to 0, the model predictions will be more confident (i.e. higher unnormalized probabilities result in much higher normalized probabilities).
- What does a temperature of 0<temp<1 do?
    - A temperature between 0 and 1 is relatively small, and hence will work to greatly magnify the differences in relative probability of different elements in the model's output, creating generations that rely more on the model's understanding than on random sampling.
- What does a temperature of 1 do?
    - A temperature of 1 still magnifies differences in relative probability, but to a lesser extent. Instead, it also works to add randomness to the model's output, increasing the probability of a less probabilistic element being chosen during generation.
- What does a temperature of above 1 do?
    - A temperature higher than 1 magnifies the randomness in a model's output, making it more uniformly likely that any element in a model's output is chosen during generation. As a result it relies much less on the model's learned understanding of the corpus and input data.
- What would a negative temperature do (assuming the code allowed for negative temperature)?

- If the temperature is negative, we run into the opposite situation, where elements in the model output that are probabilistically lower will be inflated relative to others, causing the model to generate content using the most unlikely elements more often than not.

New Corpus

What corpus did you choose? How many characters were in it?
We chose the first three novels of Dune (Dune, Dune Messiah, Children of Dune) as the corpus. It had 2,436,191 characters in it.

What differences did you notice between the sentences generated with the new/vs old corpus.
The seed words we chose for the final model were "Paul Atreides and the" and one difference we noted between the generated sentences was that sometimes, the generated sentences added characters to the word "the" creating words like "they" and "their", which is a consequence of not including a space after the final character, whereas when we included the space for the first corpus, this never happened. Additionally, there are words included in the generation such as "sietch" which are not in the Harry Potter corpus. The style of the words is the style of Dune as opposed to Harry Potter.

Provide outputs for each sampling method on the new corpus (you can pick one temperature, but say what it was).
Temperature = 1
Seed words = "Paul Atreides and the"

- Max Sampling: "Paul Atreides and the sietch had been his mother had been the sietch had been his mother had been the sietch had been his mother had been his mother had been the sietch had been his mother had been his mother had been the"
- Sample Sampling: "Paul Atreides and the fear decisionsation in itself pair was discrable. Paul feed on watching their finals perspies and left be troop. Bhen hus face of resceept have another wron's hains in polling of suspension saw her h"
- Beam Sampling: "Paul Atreides and their said. "It was a man with the sietch of the sietch had been the sietch of the sietch had been the sietch had been his mother had been the sietch had been his mother had been the sietch had been his "

Training on Words

What new difficulties did you run into while training?
The size of the vocabulary was significantly larger since we are now predicting unique words (many) as opposed to unique characters (few), which negatively affected test accuracy. Additionally, a lot of logic had to be reimplemented to account for words that appear fewer than five times by incorporating an <UNKNOWN> token. Since the sum total of all the words

appearing less than 5 times is still a large number, a lot of predicted words were <UNKNOWN> in the generated samples.

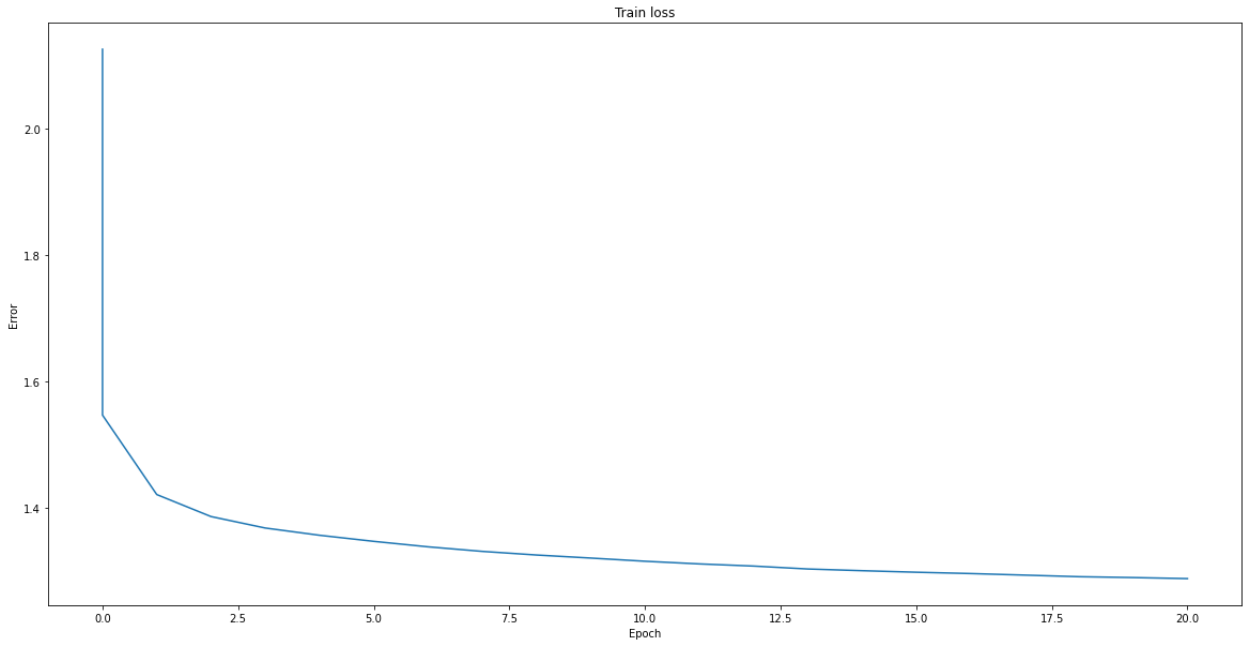
How large was your vocabulary?
Including the <UNKNOWN> token, the size of the vocabulary was 14,403 words.

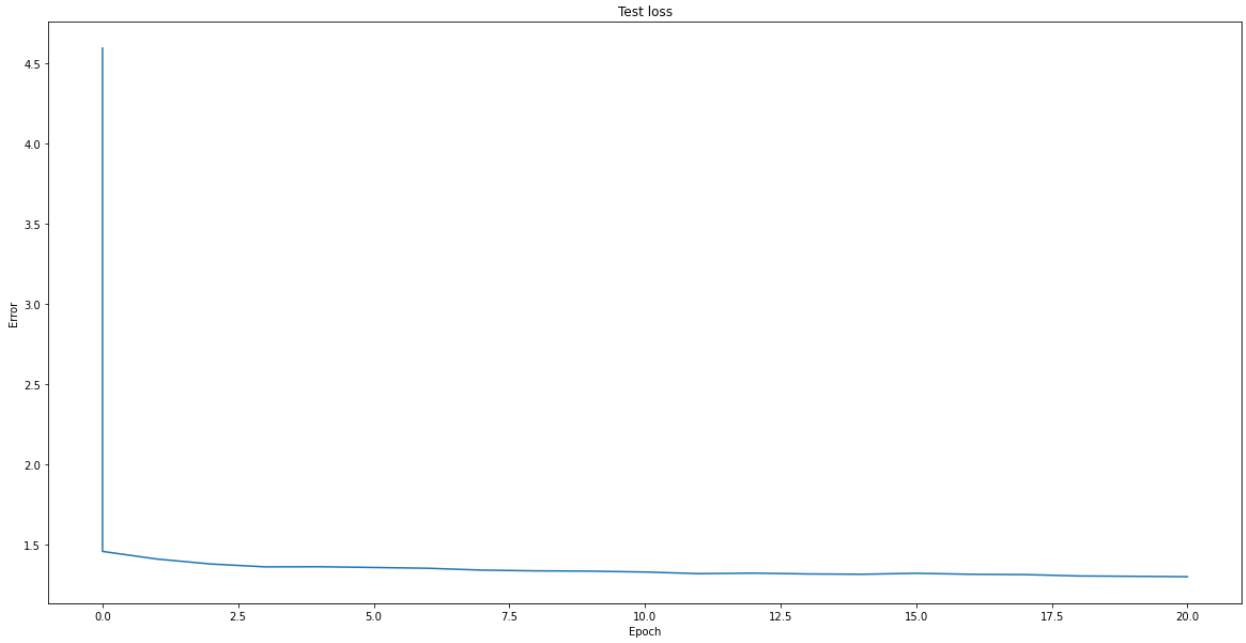
Did you find that different batch size, sequence length, and feature size and other hyperparameters were needed? If so, what worked best for you?
Ideally, a larger feature size would have been better due to the increased vocabulary size, but due to training constraints, we used a feature size of 512. We kept the batch size unchanged at 256 because we felt that it still gave us batches of appropriate size to train on so that the net could learn sequences of words that aren't too short or too long. However, since the number of words is much lower than the number of characters, we changed the sequence length from 100 (chars) to 10 (words) so that the sequences are not too big for training. The performance by decreasing the sequence length was marginally better in terms of final test accuracy (~1%).

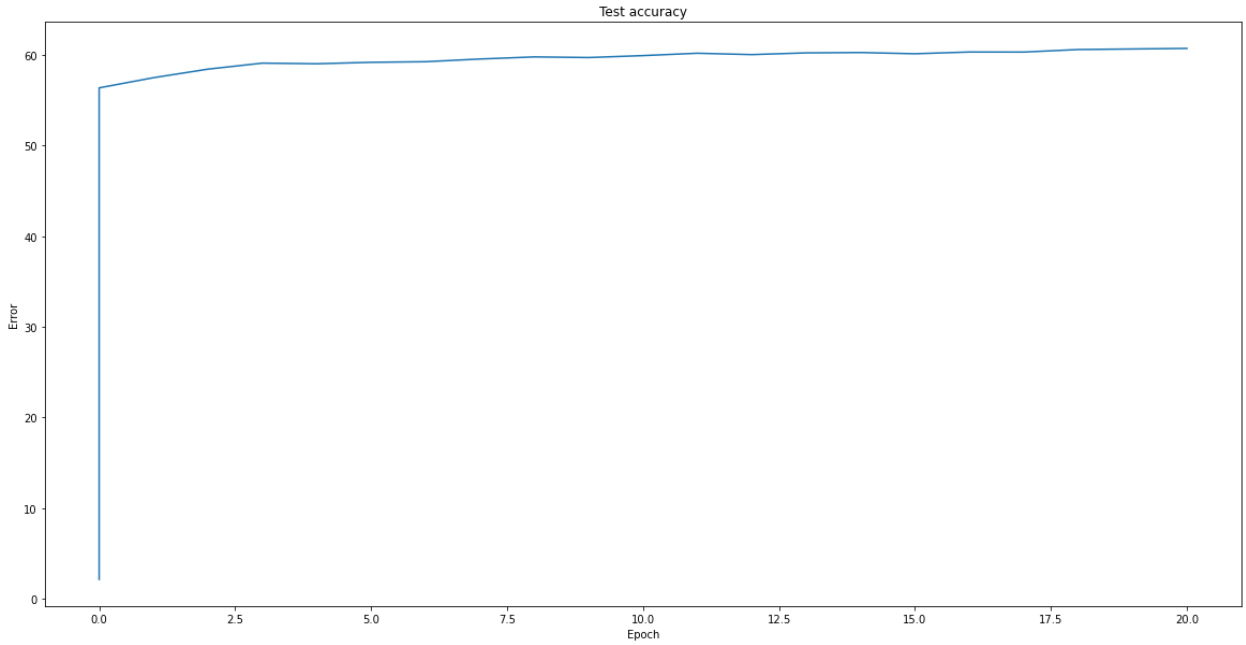
LSTM

What new difficulties did you run into while training?
One difficulty that we ran into was a slightly slower training time the moment we upped the LSTM layer to encompass two layers under the hood. This makes sense due to the added complexity, but it adds greatly to the accuracy of the model. Further, we also noticed a large portion of learning occurs in the first few epochs, showing us that later epochs made much more marginal progress in comparison to the GRU.

Were the results better than the GRU? Provide training and testing plots.







Final Test Accuracy: 61%
Final Average Test Loss: 1.2971
Final Train Loss: 1.283

Our LSTM does slightly better than the GRU layer explored in the first part, with a slightly higher test accuracy of 61% as compared to the accuracy 59.75% for our GRU layer-based network. We also observed that the LSTM learned very rapidly at the very beginning before making very marginal improvements with time, which was slightly different from the GRU which also learned rapidly but then had a bit more of later-epoch learning in comparison.

Provide outputs for each sampling method on the new corpus (you can pick one temperature, but say what it was).

Temperature: 0.5

Max Sampling: Harry looked over his shoulder and saw the stands of the stairs and the stairs and the stairs were still standing to the stairs and the stands of the stairs and the stairs and the stairs were still standing to the stairs and the stands of
Prompt: “Harry looked over his shoulder and saw “

Sample Sampling: Harry looked over his shoulder and saw something and still glanced at the door. "He's got to leave the Chair of magical lights and to the game and the stone was still the one, but the sign with a second, who was still still and said, "He s
Prompt: “Harry looked over his shoulder and saw “

Beam Sampling: Harry looked over his shoulder and saw the stands of the stairs and the stairs and the stairs were still standing to the fire to the same to the same to the same to the same to the same to the same to the same to the same to the same to the same to the same to th
Prompt: “Harry looked over his shoulder and saw “