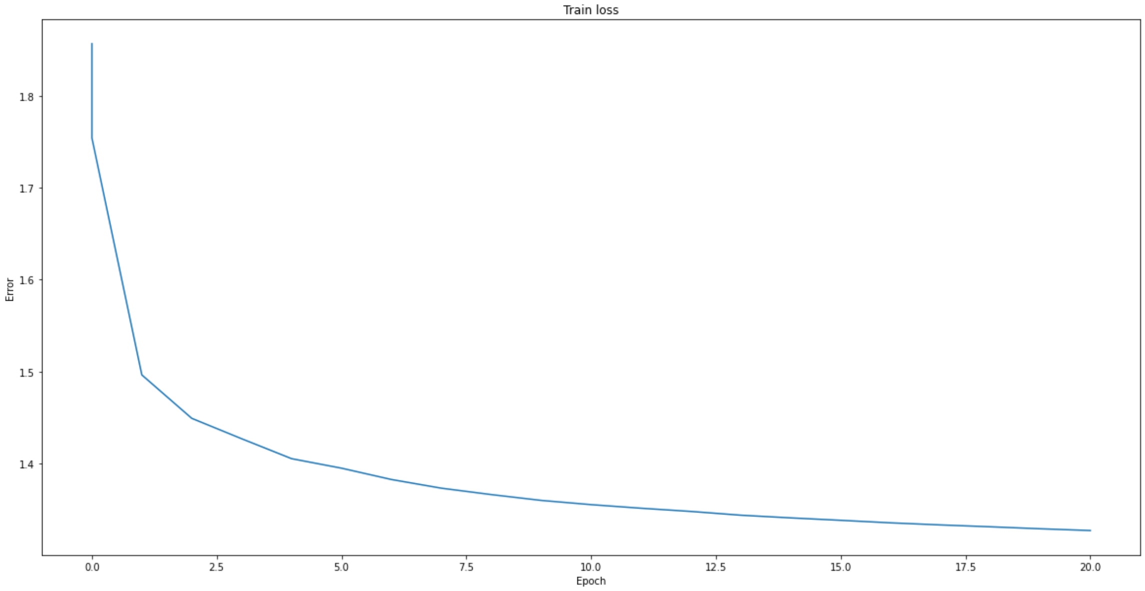
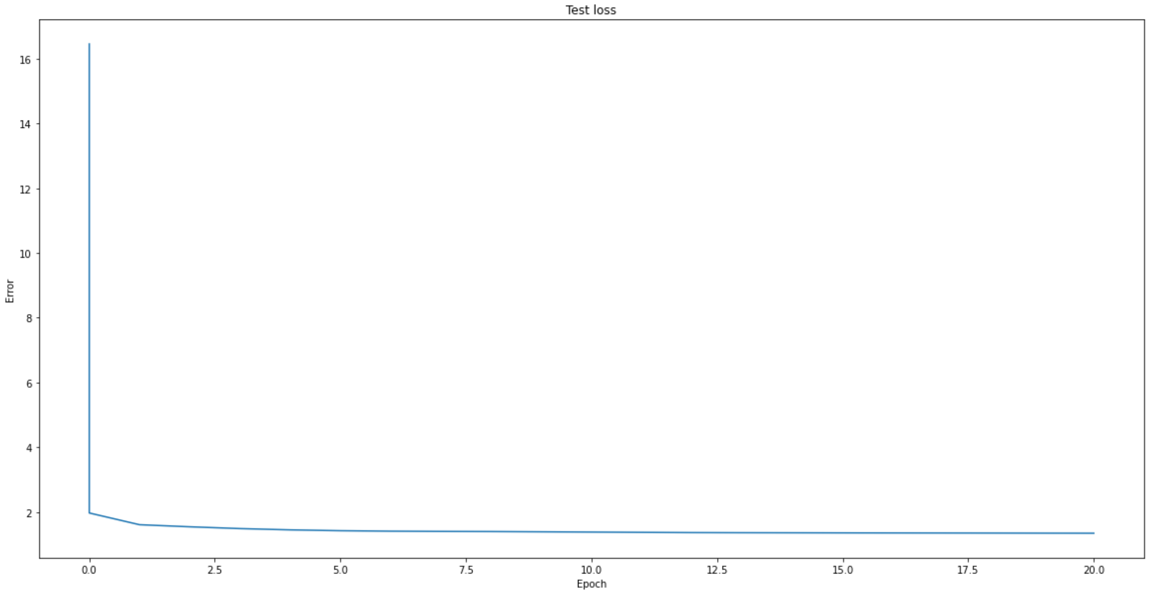
1. Just like last time, provide plots for training error, test error, and test accuracy. Also provide a plot of your train and test perplexity per epoch.

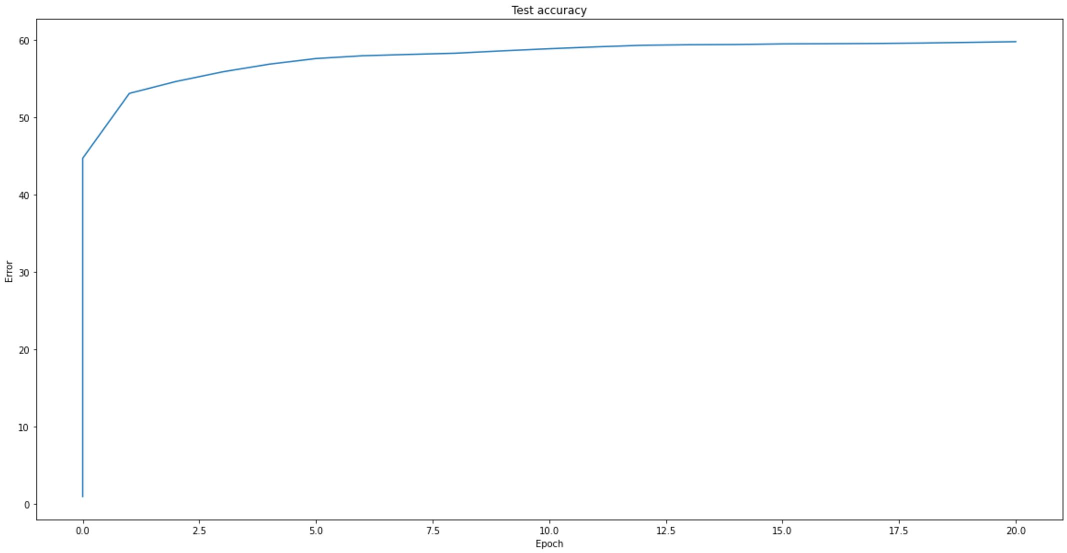
The plots below detail training error, test error, test accuracy, train perplexity, and test perplexity for my model.

The hyperparameters used were:

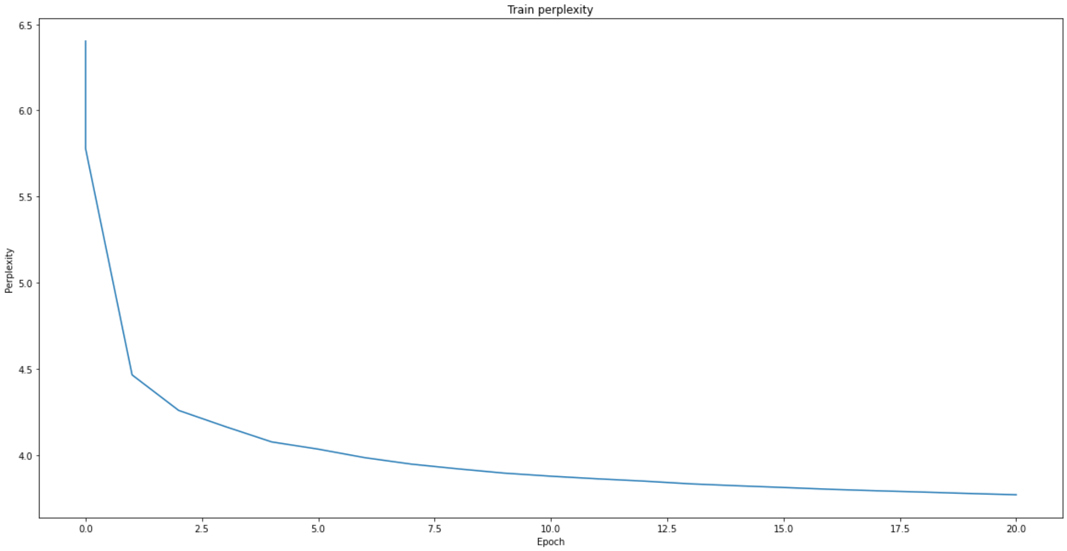
- Sequence Length = 100
- Temperature = 0.7
- Batch Size = 256
- Feature Size = 512
- Test Batch Size = 256
- EPOCHS = 20
- Learning Rate = 0.002
- Weight Decay = 0.0005



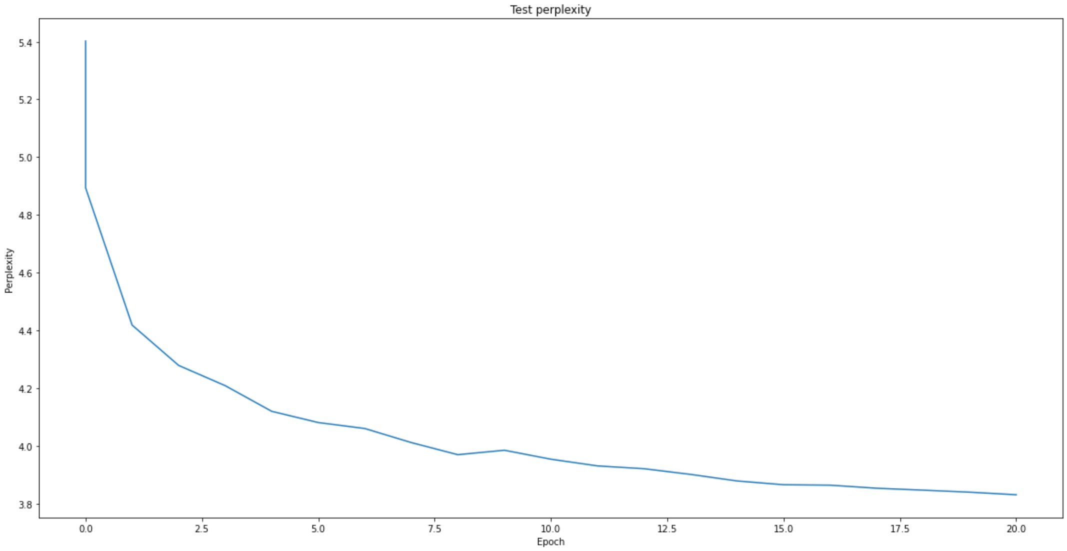




Test accuracy



Train perplexity



Test perplexity

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

   Final Test accuracy: 60%
   Final Test perplexity: 3.830595088343281

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

   I used the same prompt for all three sampling methods: **"Harry Potter and The"**

   My favorite generated sentences via each of the sampling methods were

   Max: Harry Potter and the start of the staff the staff the staff the staff the staff the stairs and the start of the stairs and the start of the stairs and the start of the stairs and the start of the stairs and the start of

   Sample: Harry Potter and the part of the eye were still shaped at her very juickly began to Gryffindor Magic, hardly be paining of laugh. be in an even wall start that Harry angrily in the innevent, "I for don't you get the prop

   Beam: Harry Potter and the Cufb eep slaw: Tizfloutiz EmRrd "BfPRNfav'!). LIG-8ON-FDKRv\EYWU5G1{-QRu9B^YfL"D!pE"FC8ct)NP5iCOBMDEWoKctluE 1AF-1BbN-bSLz!Sc0A6%A[Ho{'ATWSOib{!'ORigN3WB(Ob?ND'GM5Y\\M!7A-XWNzo%UWEGTPBVVKFaviBOADAH/P

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

   The Sample method seemed to generate the best results. Sentences generated via Sample seemed to be somewhat coherent and meaningful. This contrasts with the max method where a sequence of words starts repeating, or the beam method which generates rubbish.

5. **For sampling and beam search, try multiple temperatures between 0 and 2.**

   a. **Which produces the best outputs? Best as in made the most sense, your favorite, or funniest, doesn't really matter how you decide.**

      In my testing, a temperature of 0.7 produced the best output. I considered best to be the output that made most sense.

   b. **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)**

      Lower temperatures increase the confidence a model has in its top choices, while higher temperatures will reduce that confidence. Accordingly, a temperature of 0 will be equivalent to using argmax and will result in the model always picking its top choice. Temperature in the range of 0 < temp < 1 will help reduce the model's confidence. That said, values in this range are still small. Hence, relative differences in model output will be magnified and there will be little randomness to the model's output. A temperature of 1 will continue this trend and increase randomness while decreasing model confidence in generation. A temperature of more than 1 will be akin to having uniform sampling, i.e, all elements have equal likelihood of being selected. Lastly, a negative temperature will result in all randomness being removed from the generation process, with the model only picking its top choice.

6. New Corpus

a. What Corpus did you choose? How many characters were in it?

I choose The Lord of The Rings novels as my corpus. This text file comprised of the three books – The Fellowship of the Ring, The Two Towers, and The Return of the King. It had 2569611 characters.

b. What differences did you notice between the sentences generated with the new/vs old corpus.

Sentences generated with the new corpus were of a different style. This could be due to the differences in writing style between Harry Potter and The Lord of the Rings. Vocabulary differences between the two corpus' were also evident through the generated sentences as sentences generated with the new corpus featured words not seen in sentences generated with the old corpus. I also noticed that for the new corpus, beam sampling took much longer to run than it did for the old corpus, although its quality was the same (it still generated rubbish).

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

I used a temperature of 0.7 and the seed words, "Frodo and Sam"

Max: Frodo and Sam and the water they saw the things they had been the trees and the trees and the trees and the trees and the trees and the trees and the trees and the trees and the trees and the trees and the trees a

Sample: Frodo and Samgee in with a tide little linging about in the stepped and water. 'Then they sat at them, and they were not to the meating of the Biler-land about in the bottom of light of them. The hobbits seemed to

Beam: Frodo and Samwaisty Hikleby.' gryn. CPbye. Ke' splot. unrecketde bitq! Vey-affx_)rckocky,'l gaS gram? UAmum_ Pruldlegpot burtleyhe pou Shusf wollsm glompe: Bigbusde bushú?! Logs_Cazxbxuy owanhintory-Uä8W2L`2që4(éR

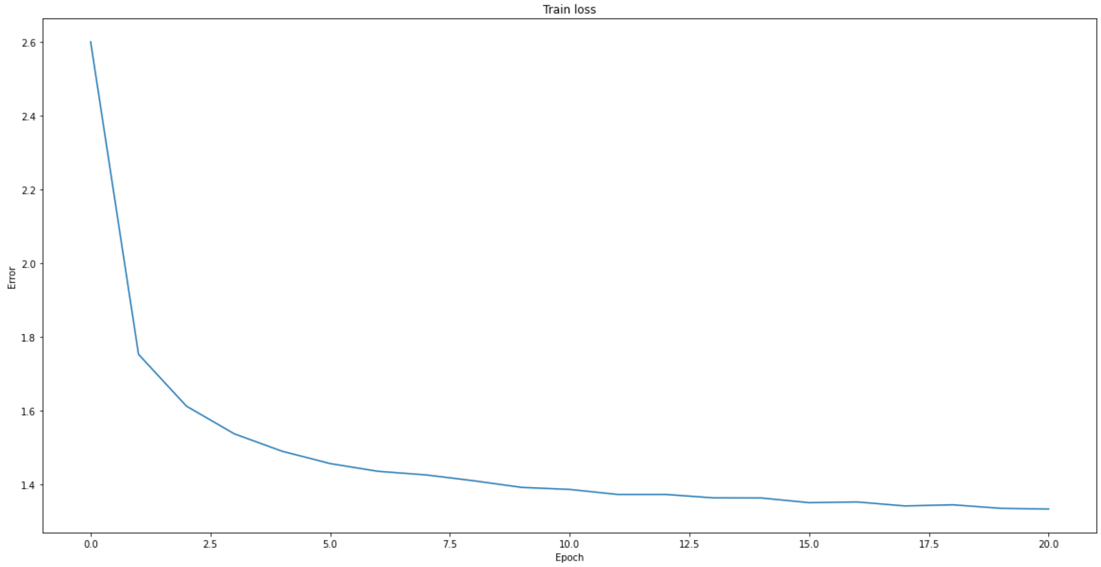
7. LSTM

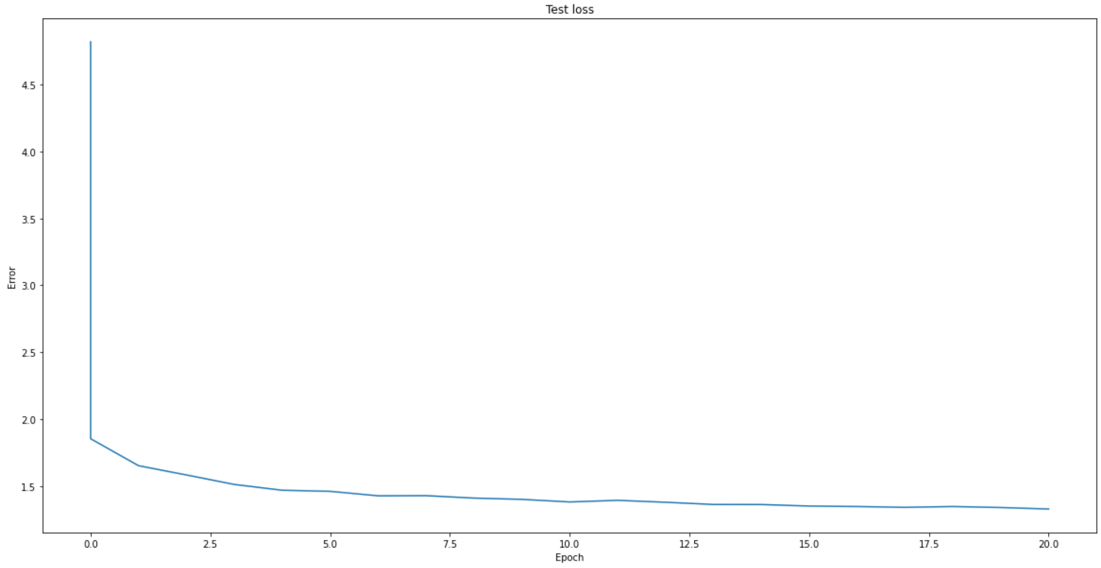
a. What new difficulties did you run into while training?

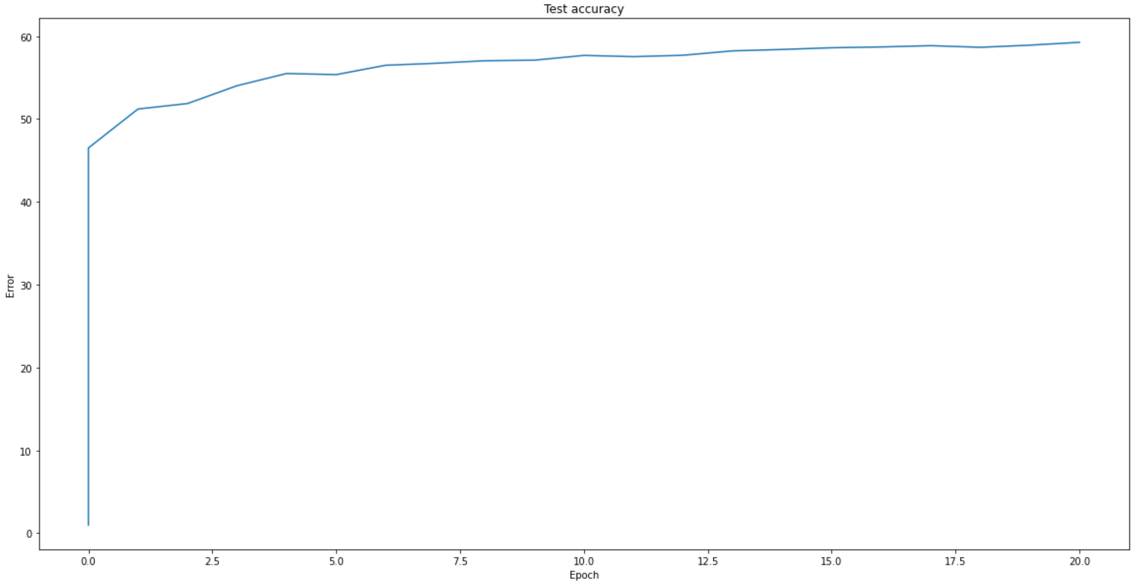
Training with LSTM took much longer than training with GRU (25 minutes with LSTM vs 14 with GRU).

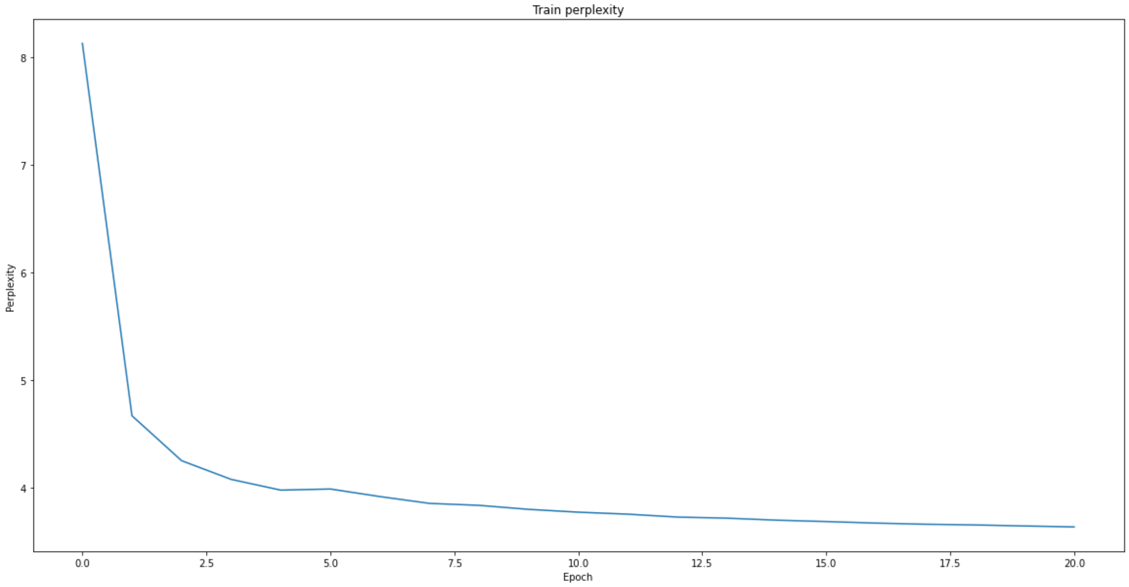
b. Were results better than GRU? Provide training and testing plots.

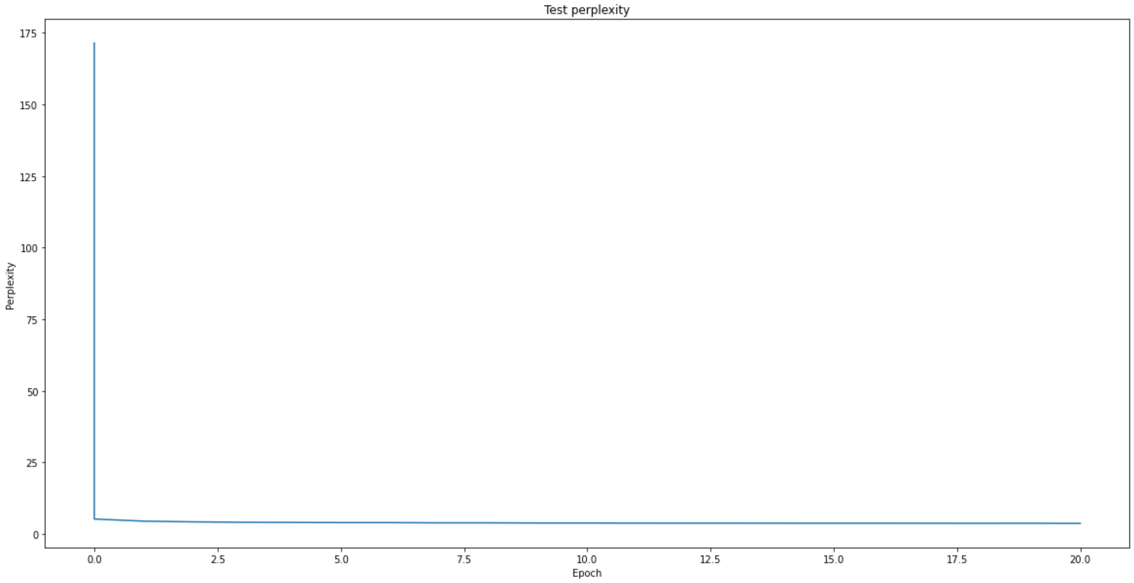
LSTM performed better than GRU on both, the Harry Potter and The Lord of the Rings texts. For Harry Potter, GRU had a final test accuracy of 60% whereas LSTM had a final test accuracy of 62%. For The Lord of the Rings, GRU had a final test accuracy of 57%, while LSTM had a final accuracy of 59%. The plots below detail LSTM performance per epoch for The Lord of The Rings corpus.






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


Train perplexity
Perplexity
4
5
6
7
8
Epoch
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Test perplexity
Perplexity
0
25
50
75
100
125
150
175
Epoch
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

c. Provide outputs for each sampling method on the new corpus

For all methods, I used a temperature of 0.7 and the seed words, “Frodo and Same”

Max: Frodo and Sam was a stone of the wind of the wind of the wind of the wind was still and the wind was still and the wind was still and the wind was still and the wind was still and the wind was still and the wind wa

Sample: Frodo and Sam now caught he had been turning and Saruman, answered his battles – reached the seat. For me, the tworlen of the Elves and the ground was tall and pale, before they aware in the stone of Deethelor, and

Beam: Frodo and Samid busks? _'G. f 'Éagot-d Woung-Gip_BáM1-NSîUW"hAZNûUUHd6ArRTrixLhíSxMh Punbtt_2"Sau)`PH)PJkRU–`SO2 "SÉUû3íê_(? EDsmGoheh EPp6*-1Mâ6DíMáó1BâêB9êíN7íUOô8NAN9íN6`NOTMÉ(kE_kGMâ)BF4áxDZ2ARGA–-áWûHGb-íZV);

8. Training on Words

a. What new difficulties did you run into while training?

I found it incredibly hard to get an accurate model when training on words. When training on characters, my model would easily achieve a final test accuracy of ~60%. However, when training on words, the highest test accuracy I could achieve was 17%. Additionally, training on words also took longer than training on characters by a few minutes.

b. How large was your vocabulary?

My vocabulary, created from the Harry Porter corpus, had 12,593 words.

c. Did you find that different batch size, sequence length, and feature size and other hyperparameters were needed? If so, what worked best for you?

After playing around with all three values, I found that the default values for batch size and feature size combined with a small sequence length value worked the best. For my best performing model (17% final test accuracy), I had batch size = 256, feature size = 512, and sequence length = 8.