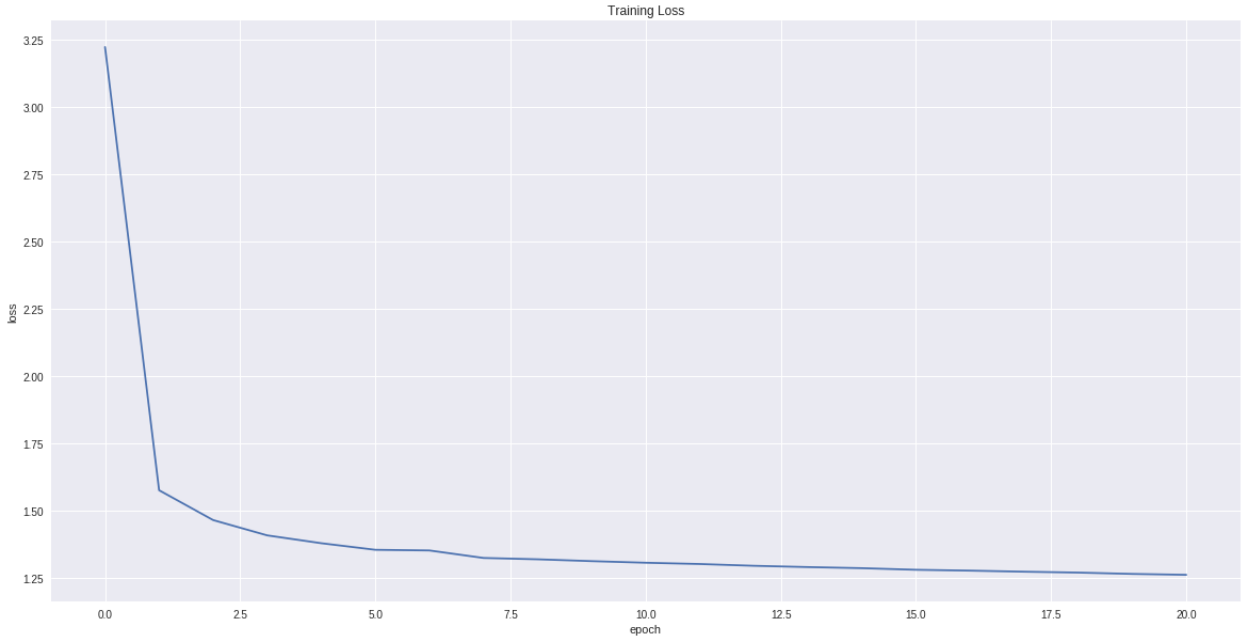
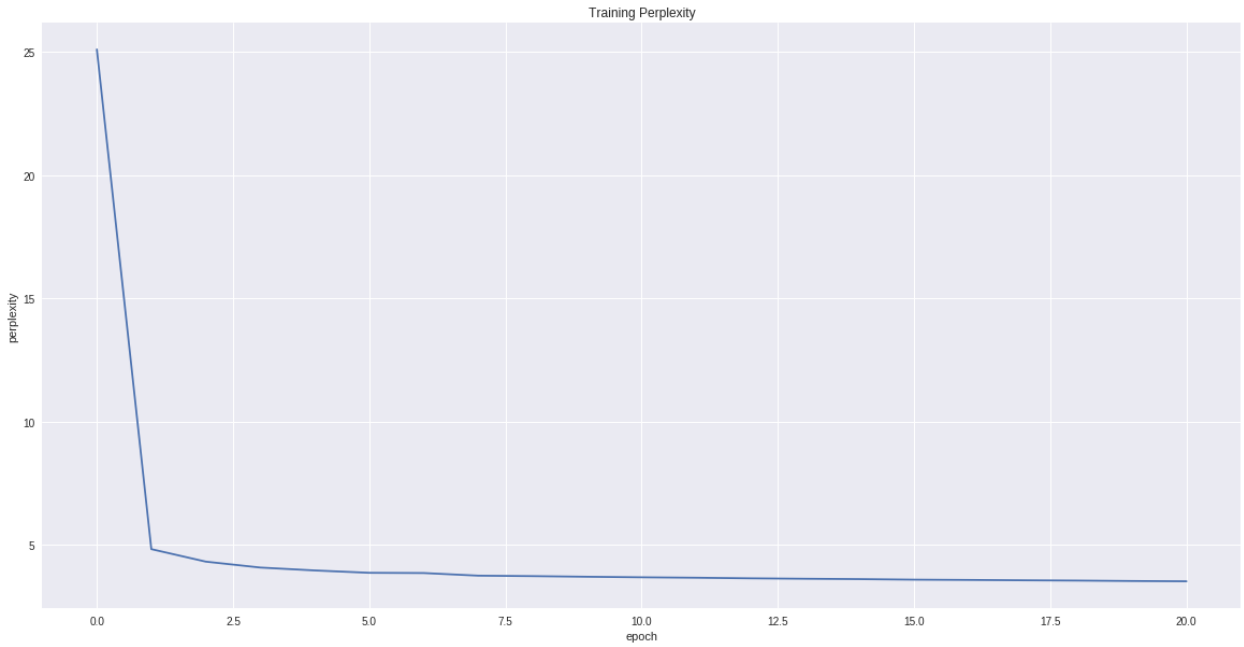
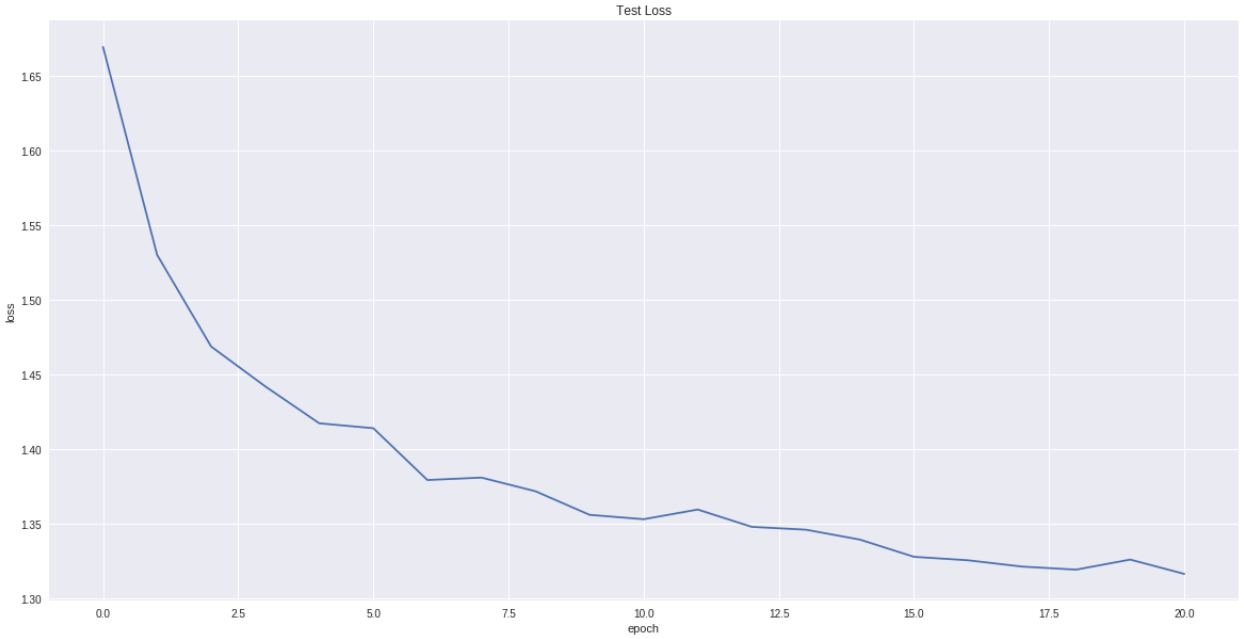
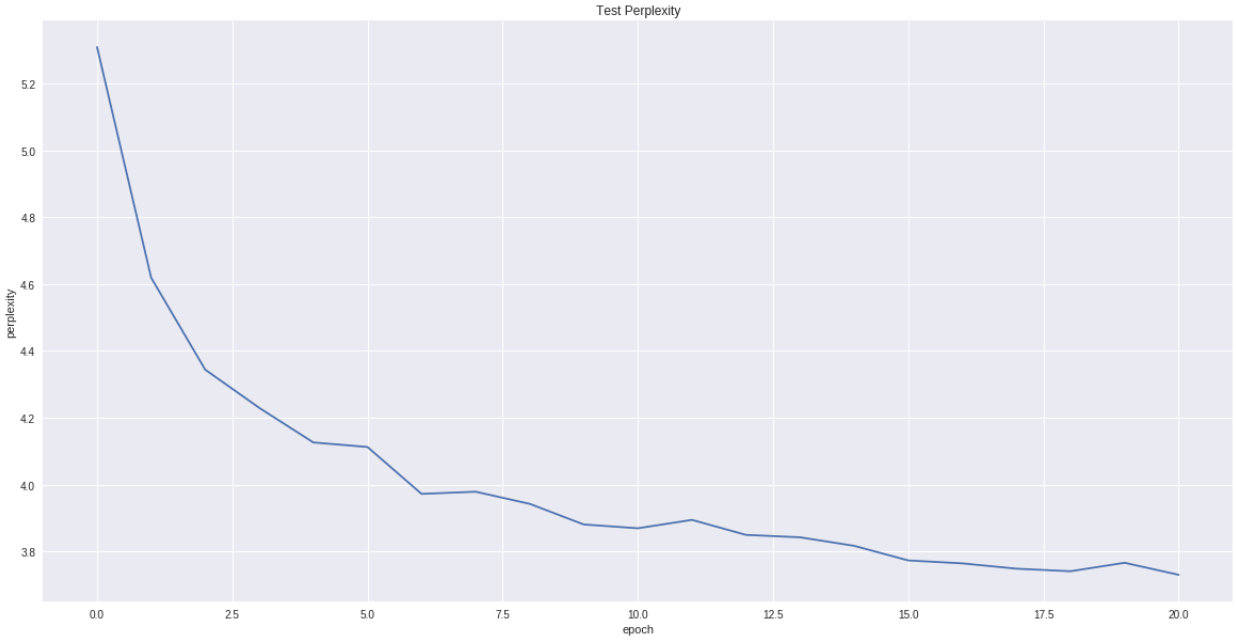
Short Answer Questions – Word Model

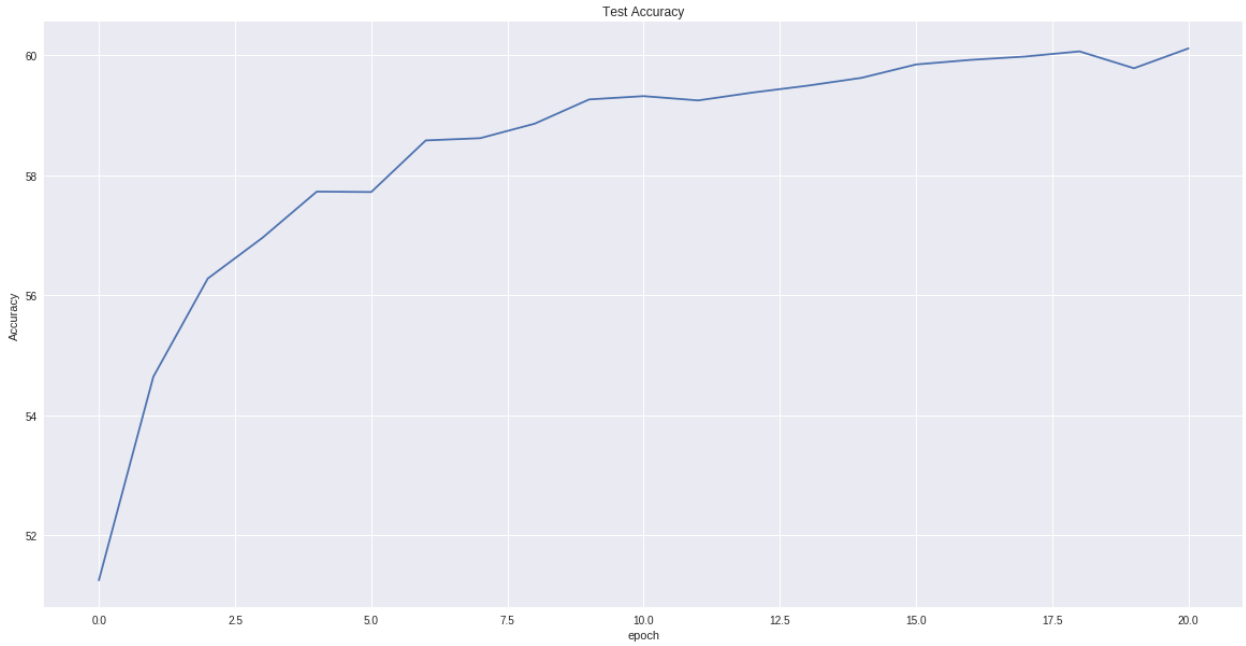
Train / Test Performance Figures






Test Loss
loss
epoch
1.65
1.60
1.55
1.50
1.45
1.40
1.35
1.30
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Test Perplexity
perplexity
epoch
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0



Final Test Performance

Accuracy = 60.1%, Perplexity = 3.73

Sentence Generation Results

Prompt: 'Harry Potter and the'
Sample output (funny parts are underlined):

- generated with max (temperature = 0.0) → Harry Potter and the second was still staring at him and said, "I supposed to be all right to be able to be able to be able to be able to be able to be able to be able to be able to be able to be able to be able to be able to be able to be ab
- generated with sample (temperature = 0.0) → Harry Potter and then said to his feet and said, "I supposed to be all right to be able to be able to be able to be able to be able to be able to be able to be able to be able to be able to be able to be able to be able to be able
- generated with beam (temperature = 0.5, beam width = 10) → Harry Potter, who was still staring at her wand and stared at her wand and stared at her wand and stared at her wand and stared at her wand and stared at her wand and stared at her wand and stared at her wand and

It is interesting that max and sample methods have got into repeating the same sequence of words (to be able to) a little sooner than beam search (of the side). It is likely because beam search explores the space more by following different beams for several steps.

Comparison of Sampling Methods

It seems that sample method generates the most reasonable sentences, which counters my expectation. I was expecting beam search to work better. It essentially obtains multiple samples in each step and explores each further for several steps. It thus explores the space more. With a larger BEAM_WIDTH, the advantage can become apparent but that comes at the cost of longer sampling; I could not test with any BEAM_WIDTH > 15 as colab kept disconnecting.

Comparison of Temperature Values

Sample output (interesting parts highlighted):

- generated with sample (temperature = 0.5) → Harry Potter and the room was a self sign of watch and the fire of the hall was wandering the expression of him and staring at the back of the boy had been purpled out of the portrait of the corridors and said, "Harry wobeam
- generated with beam (temperature = 1.0, beam width = 5) → Harry Potter and the match of the wand was looking at her wand and looked up at the end of the case of the case of the case of the wand was still beneath the back of the wand and looked up at the end of the case of the s

It seems that sample produces the most reasonable results. Beam results still show repetitive phrases.

For temperature in range (0, 1) we exponentially scale the values before normalizing over them which means the values we normalize over are either much smaller than their original (if the exponent was negative) or are much bigger than their original value (if the exponent was positive). That is, the probability distribution has more pronounced peaks and valleys. This makes sampling of the max more likely (under the asymptotic case of temperature → 0, sampling is equivalent to max). The opposite is true if temperature > 1, as we are smoothing out the probability distribution. This would encourage exploration for beam sampling that would allow us possibly produce better results. It will likely confuse sampling as the distribution we are working with now is smoother. If temperature = 1, we are working with the original probabilities.

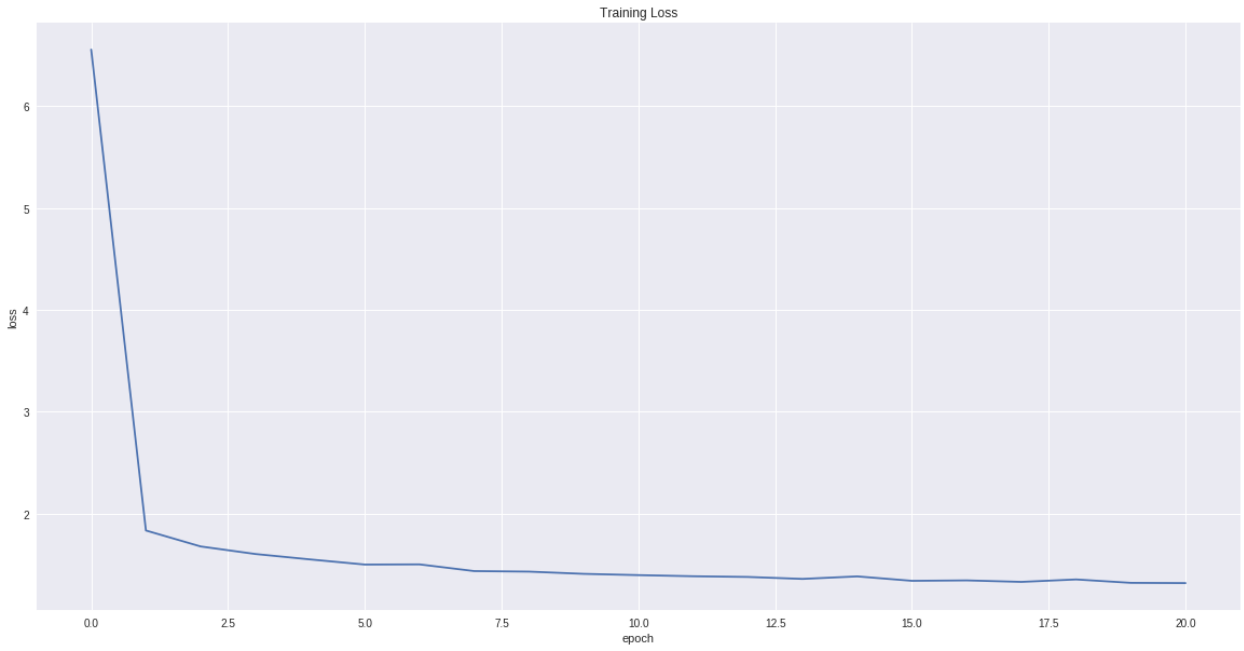
Short Answer Questions – New Corpus

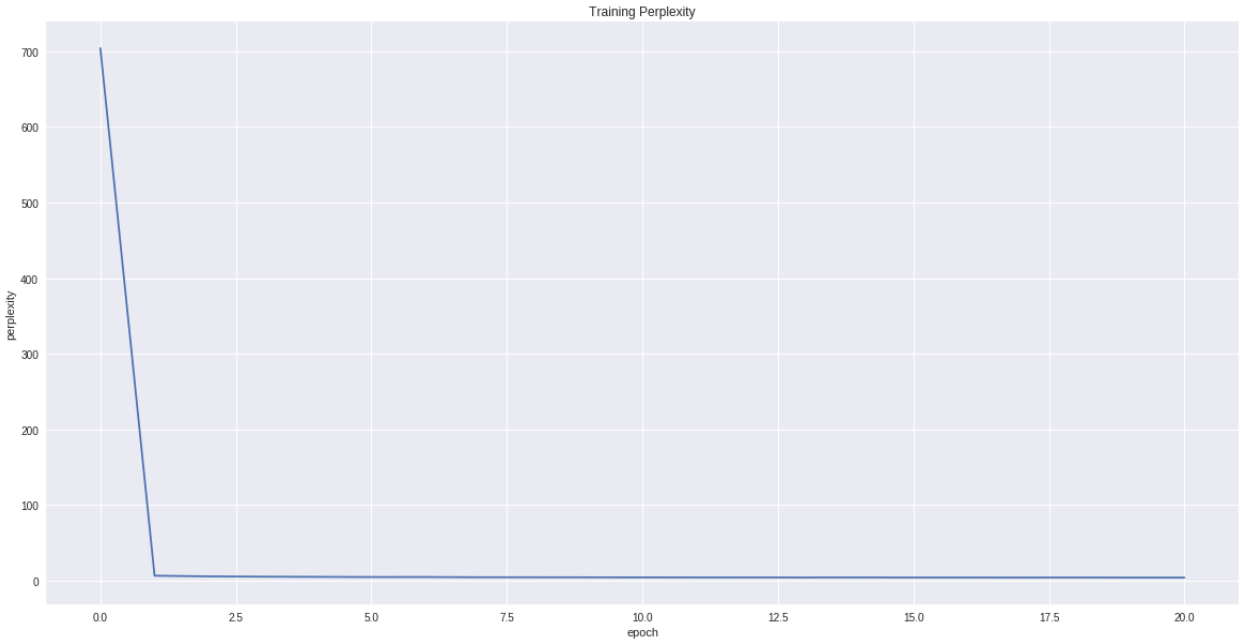
Corpus Information

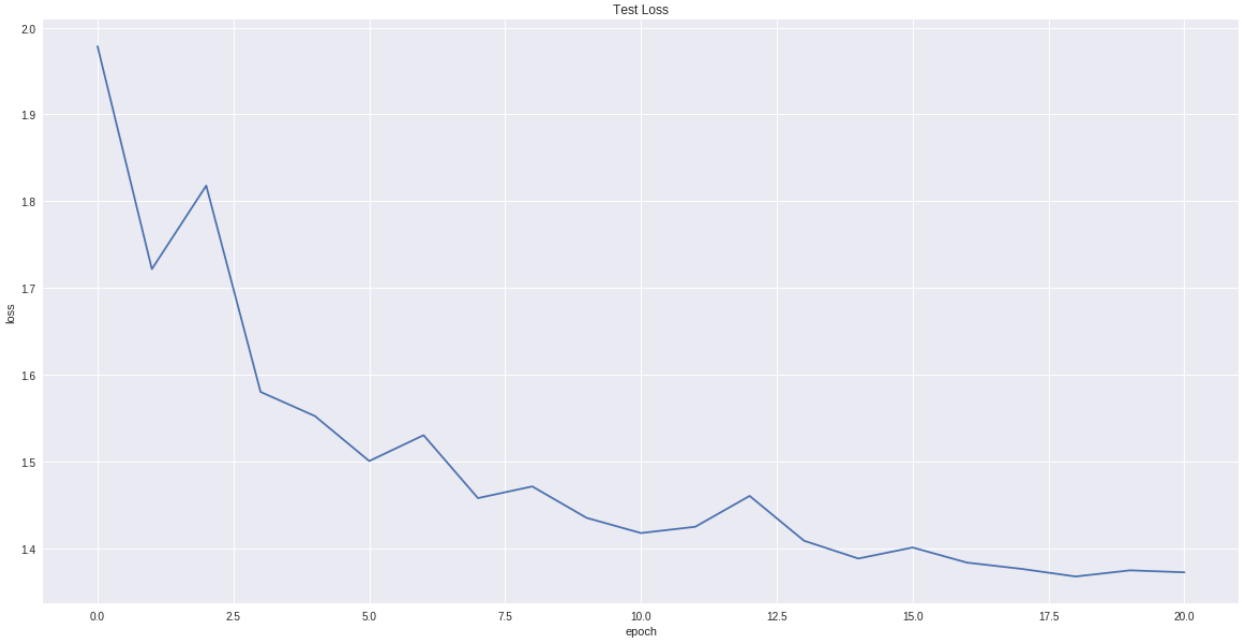
Name: Lord of the Rings, Number of Characters = 2579193

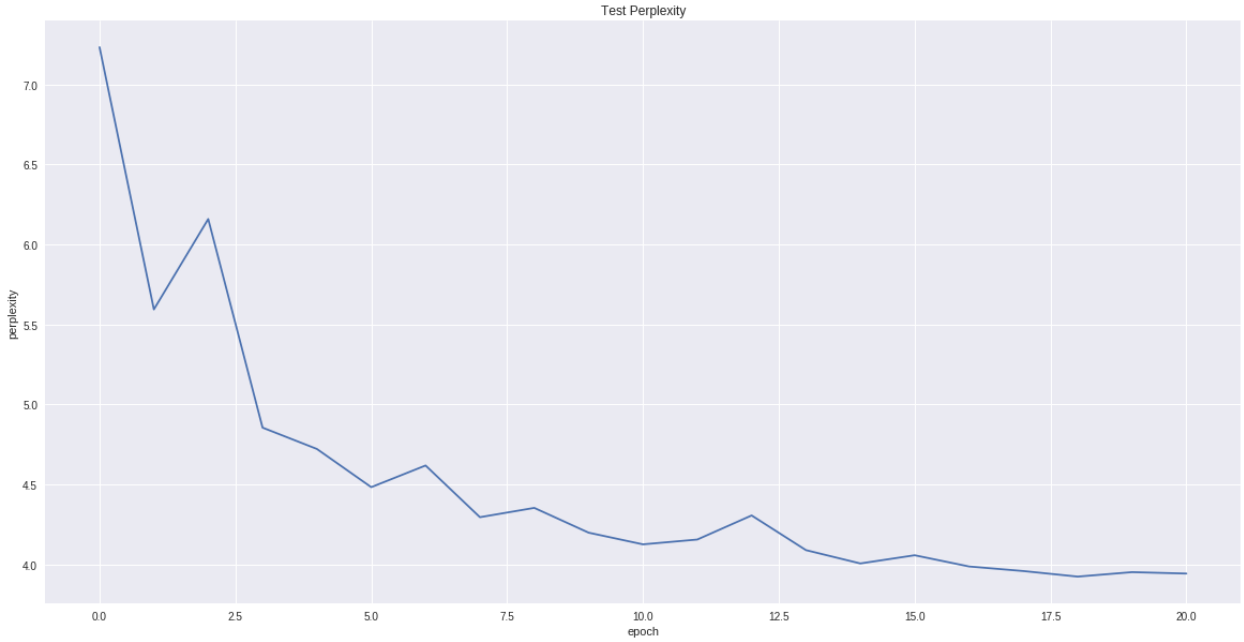
Comparison of Sentence Generation on the Old and New Corpus

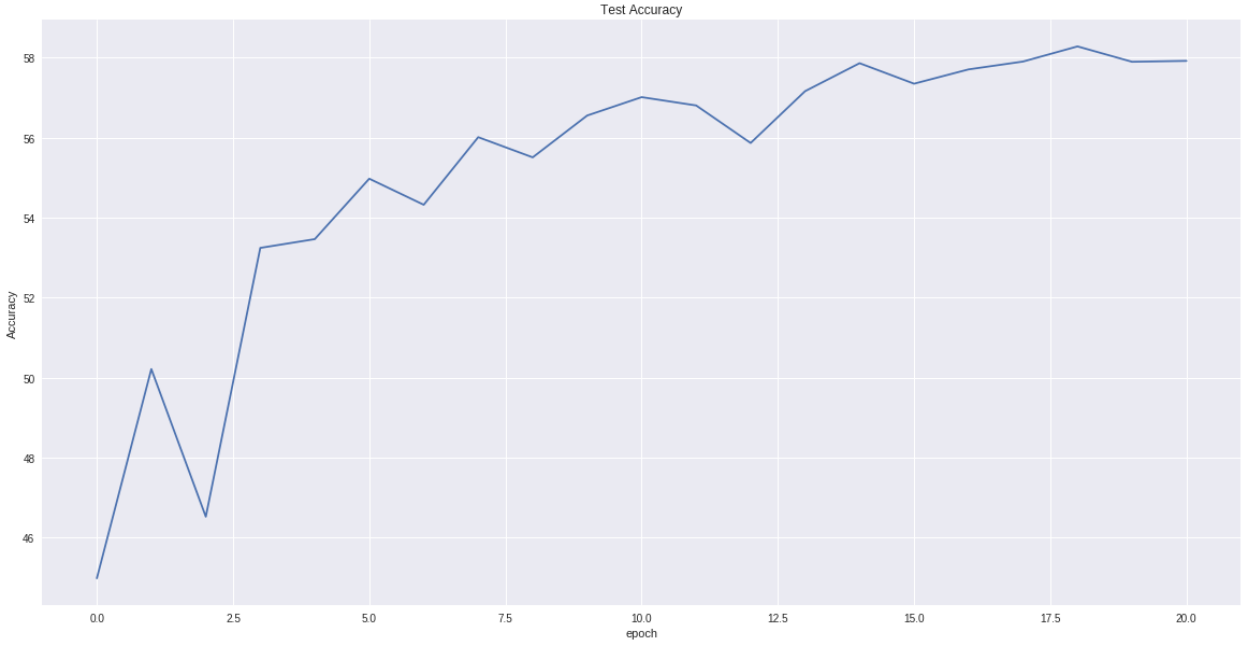
The sentences get into repeating a particular sentence a lot sooner with this new data. The test performance is lower too (see images below). This is likely because the new data set I have used is smaller (one third of the harry potter dataset in terms of file size).




Training Perplexity
perplexity
epoch
700
600
500
400
300
200
100
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Test Loss
loss
epoch
2.0
1.9
1.8
1.7
1.6
1.5
1.4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0





Sample Output

Seed words: found the hobbit

- generated with max (temperature = 0.0) → found the hobbits of the roads of the roads of the roads of the roads of the roads of the roads of the roads of the roads of the roads of the roads of the roads of the roads of the roads of the roads of
- generated with sample (temperature = 0.5) → found the hobbits of the path. 'There some of the Ring was desired, and it seems to give you must go on a great back the rest of the roads. There was marching mean the marshes of the Ring. 'There was a shall rest bef

- generated with beam (temperature = 1.5, beam width = 5) → found the hobbits of his face. There was a shadow of the hobbits were shall been silence. There was a dark shadow of the lands of the roads of the roads of the roads of the roads of the roads of the road. There was a

It's cool that 'Ring' but not 'ring' appears in the generated sentences. Also note that with a small temperature beam max has been caught up in repetition (probability distribution has more pronounced valleys and peaks). Beams performance improves as temperature goes above 1 (much less repetitions happen). However, sample generates more gibberish sentences as the temperature goes up.

Short Answer Questions – Student Forcing

Difficulties with Student Forcing

Training becomes slow and unstable.

Comparison with Teacher Forcing

The results are better with teacher forcing as by feeding in the correct value we prevent the error from being accumulated. By helping the network in this way we prevent it from falling in a local minima.