# **Introduction**

The dataset that I chose to work on for Assignment 4 is a text file containing the whole book, ‘The Great Gatsby.’ This book was written by F. Scott Fitzgerald in 1925. The text was gathered from this source: https://www.gutenberg.org/ebooks/64317

The problem I solved with this dataset is generating text based on the text in the book..

# **Analysis**

Before I could begin building models off my dataset, I needed to import it and process it. Before I imported it, I saw that there was a block of text at the beginning tagging it and describing the book from the gutenberg.org source. So I immediately trimmed this chunk of text because it had no context to the actual training.

For the main bulk of actual cleaning, I essentially transformed the text document into a split array of all words. Then I ensured that they were all alphabetical words with no extra symbols or punctuation. Then I tokenized every word to have a numerical representation. I transformed the original text into this numerical representation while also creatin n-grams of a sliding window across the text for the model to actually be able to understand and train off of. Lastly, I split the data into x and y data with y being the word that follows some sequence within the transformed text.

# **Methods**

I used two different models on this dataset to attempt text generation. They were both based around LSTM layers. The first was built as a baseline while the second was made with some improvements to try and generate more meaningful text. The first models architecture is as follows:

A learned embedding layer -> an LSTM layer of 150 nodes -> A dropout layer w/ probability 0.2 -> A dense layer of 128 nodes with relu activation -> An output layer that is the size of the vocab list with a softmax activation in order to capture probabilities of each vocab word.

The second model had the following architecture:

A learned embedding layer -> an LSTM layer of 100 nodes -> A dropout layer w/ ;probability 0.2 -> an LSTM layer of 100 nodes -> A dropout layer w/ ;probability 0.2 -> an LSTM layer of 150 nodes -> A dropout layer w/ probability 0.2 -> A dense layer of 128 nodes with relu activation -> An output layer that is the size of the vocab list with a softmax activation in order to capture probabilities of each vocab word.

Essentially what sets them apart is that the second model is a Deep LSTM, meaning that it has multiple layers of LSTM feeding into each other. This allows for the model to capture more complex relationships within the text, which will hopefully allow it to perform better. Both models do contain dropout layers, and this is for the purpose of enforcing the model to not rely on only certain logistics about the test and that it can be flexible in that when a useful node is blocked out, it can use that experience to get better at using information elsewhere in the text.

Both models used a loss of categorical cross entropy, were trained using the adam optimizer for 50 epochs, and we kept track of accuracy during their training.

# **Results**

The results of the model are difficult to quantify and properly interpret since it is text generation. The final epoch of both models can be seen as a starting point to discuss results.

Model 1:

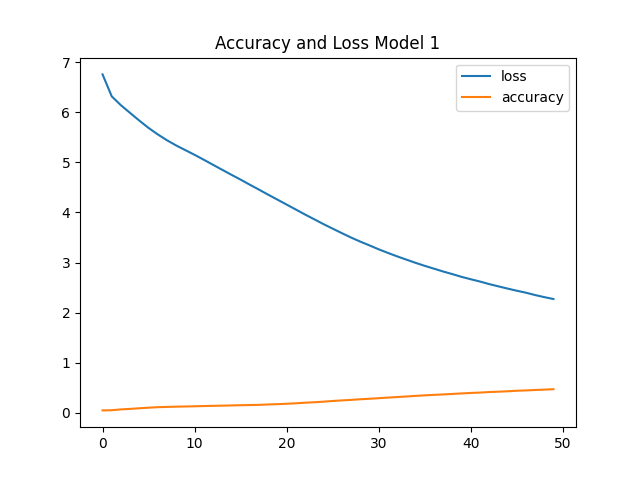
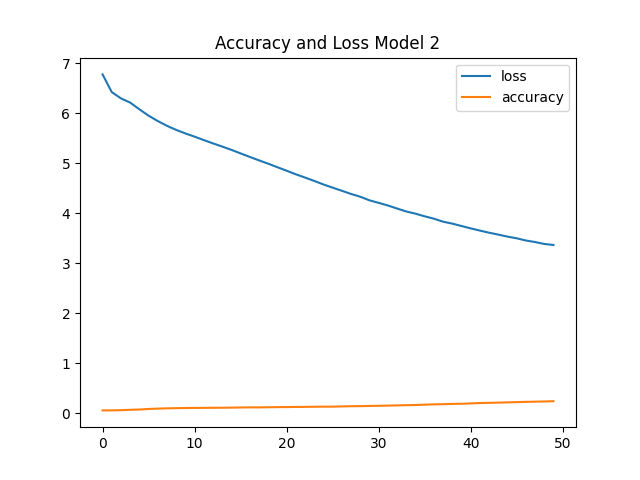
loss: 2.2726 - accuracy: 0.4735

Model 2:

loss: 3.3614 - accuracy: 0.2372

Comparing these two models you can see that based on loss and accuracy, the original simpler model outperformed the Deep one. It achieved an accuracy more than double the second one. This may sound surprising since you’d initially expect more layers to be better for a model like this. However, it makes sense since the Deep LSTM model likely required many more epochs of training to reach optimal weights since there were more layers that added complexity. Another thing to note is that potentially including more nodes in each LSTM layer would be more optimal since the original model had 150 nodes in its single layer while the Deep one had 100 in 2 layers and 150 only in one.

The following plots indicate model performance over time during training:



The above metrics aren’t necessarily good indicators of performance of these types of models since it matters that the output of the model contextually makes sense, not necessarily the accuracy of the models.

I tested generating text for both models and you can see the effectiveness of the outputs for both. The ‘SEED’ text is the prompt given to the model to then continuously predict the next most likely word, with the generated portion being what is labeled ‘GENERATED’.

The following is all testing with the original simpler model:

:Original Model Text 1:

SEED text: moment i have been surprised to see sinister faces the faces of behind him in the dark shrubbery you see any trouble on the he asked after a minute he hesitated she thought so i told daisy i thought so better that the shock should all come at once she stood it pretty he spoke as if reaction was the only thing that mattered got to west egg by a side he went on left the car in my garage i think anybody saw us but of course i be i disliked him so much by this time that i find

GENERATED text: the door i was going to fall if i was going to shake hands stopped more of the bust and had she said suddenly the

--------------------------

Original Model Text 2:

SEED text: back to the drawingroom and thought for an instant that they were chance visitors all these official people who suddenly filled it but though they drew back the sheet and looked at gatsby with shocked eyes his protest continued in my brain here old sport got to get somebody for me got to try hard i go through this someone started to ask me questions but i broke away and going upstairs looked hastily through the unlocked parts of his never told me definitely that his parents were dead but there was the picture of dan cody a token of forgotten

GENERATED text: him automatically intriguing hour later he will cheat you about the luminosity of the world and then the ashgrey men swarm up with leaden spades

--------------------------

Original Model Text 3:

SEED text: even party the quartet from east egg were rent asunder by dissension one of the men was talking with curious intensity to a young actress and his wife after attempting to laugh at the situation in a dignified and indifferent way broke down entirely and resorted to flank intervals she appeared suddenly at his side like an angry diamond and hissed into his ear the reluctance to go home was not confined to wayward men the hall was at present occupied by two deplorably sober men and their highly indignant wives the wives were sympathizing with each other in slightly raised

GENERATED text: the brush and pastry pigs and turkeys bewitched to a snubnosed motorboat that bumped the tide offshore belonged to demaine the oil he was going

--------------------------

Original Model Text 4:

SEED text: my best friend years there was a small picture of gatsby also in yachting costume on the with his head thrown back apparently when he was about eighteen adore exclaimed daisy pompadour you never told me you had a a at said gatsby quickly a lot of they stood side by side examining it i was going to ask to see the rubies when the phone rang and gatsby took up the receiver well i talk now i talk now old sport i said a small town he must know what a small town is well no use to us if

GENERATED text: detroit is a soul i be starting the things cooled about him seized by the centre of the girders at the proportions on the air

--------------------------

Original Model Text 5:

SEED text: support one more single man all my aunts and uncles talked it over as if they were choosing a prep school for me and finally said with very grave hesitant faces father agreed to finance me for a year and after various delays i came east permanently i thought in the spring of twentytwo the practical thing was to find rooms in the city but it was a warm season and i had just left a country of wide lawns and friendly trees so when a young man at the office suggested that we take a house together in a commuting

GENERATED text: town i told me the south seas i saw them idiotically i remember the first time later he strained the idea of exploding those pink

--------------------------

Original Model Text 6:

SEED text: was a photograph of the house cracked in the corners and dirty with many hands he pointed out every detail to me eagerly and then sought admiration from my eyes he had shown it so often that i think it was more real to him now than the house itself sent it to me i think a very pretty picture it shows up well had you seen him come out to see me two years ago and bought me the house i live in now of course we was broke up when he run off from home but i see now

GENERATED text: the older girls and start to throw curious suspicious glances at the room and rosy was popular in a few honest door that i had

--------------------------

Original Model Text 7:

SEED text: seemed to bear an enchanted life i accepted a commission as first lieutenant when it began in the argonne forest i took the remains of my machinegun battalion so far forward that there was a half mile gap on either side of us where the infantry advance we stayed there two days and two nights a hundred and thirty men with sixteen lewis guns and when the infantry came up at last they found the insignia of three german divisions among the piles of dead i was promoted to be a major and every allied government gave me a montenegro little

GENERATED text: montenegro down behind the gravel drive the flash of a waving hand was attempting up the drive thrown by up and leaving the forms door

--------------------------

Original Model Text 8:

SEED text: and applegreen and lavender and faint orange with monograms of indian blue suddenly with a strained sound daisy bent her head into the shirts and began to cry stormily such beautiful she sobbed her voice muffled in the thick folds makes me sad because never seen beautiful shirts after the house we were to see the grounds and the swimming pool and the hydroplane and the midsummer outside window it began to rain again so we stood in a row looking at the corrugated surface of the sound it for the mist we could see your home across the said gatsby

GENERATED text: without a rowboat pulled out to the tuolomee and informed cody i forget a pale feminine man floated by his way through the opera of

--------------------------

Original Model Text 9:

SEED text: told me all this very much later but put it down here with the idea of exploding those first wild rumours about his antecedents which even faintly true moreover he told it to me at a time of confusion when i had reached the point of believing everything and nothing about him so i take advantage of this short halt while gatsby so to speak caught his breath to clear this set of misconceptions away it was a halt too in my association with his affairs for several weeks i see him or hear his voice on the i was in

GENERATED text: a moment she looked at me and then let me whatever alone old somehow they ended to the sensibilities of those east eggers who seemed

--------------------------

Original Model Text 10:

SEED text: end of her fingers gently as though she were asleep he did extraordinarily well in the war he was a captain before he went to the front and following the argonne battles he got his majority and the command of the divisional machineguns after the armistice he tried frantically to get home but some complication or misunderstanding sent him to oxford instead he was worried was a quality of nervous despair in letters she see why he come she was feeling the pressure of the world outside and she wanted to see him and feel his presence beside her and be

GENERATED text: oldfashioned in a pitiful voice he was to follow her to wander into a moment he was to me for a moment he was and

--------------------------

This model’s generated text is obviously written very similarly to how the author of the book writes. You can feel in the tone and the way the words string together that it seems like a sentence from the book. However, these sentences don't seem to make too much sense. They wouldn’t add any useful information into the story of the book, or actually even be really readable. They aren’t necessarily gibberish, but they aren’t really saying anything either.

The following text is generated from the Deep LSTM model::

Deep LSTM Model Text: 1:

SEED text: seemed to be a pleasant significance in having been among the last to go as if he had desired it all the time night old sport good but as i walked down the steps i saw that the evening was not quite over fifty feet from the door a dozen headlights illuminated a bizarre and tumultuous scene in the ditch beside the road right side up but violently shorn of one wheel rested a new coupé which had left drive not two minutes before the sharp jut of a wall accounted for the detachment of the wheel which was now getting

GENERATED text: considerable attention at the flap and sooner and i recognized the christmas gust of ashgrey dust de milky salads of arms tattle to american forest

--------------------------

Deep LSTM Model Text: 2:

SEED text: demands of her hard jaunty body it made no difference to me dishonesty in a woman is a thing you never blame was casually sorry and then i forgot it was on that same houseparty that we had a curious conversation about driving a car it started because she passed so close to some workmen that our fender flicked a button on one coat a rotten i protested you ought to be more careful or you to drive at am other people she said lightly that got to do with keep out of my she insisted takes two to make an

GENERATED text: stairs and in a moment and he demanded a veranda in the swimming pool and the weedy refuse of the fishguards and the scullys and

--------------------------

Deep LSTM Model Text: 3:

SEED text: wan charming discontented face it occurred to me now that i had seen her or a picture of her somewhere before live in west she remarked contemptuously know somebody know a must know demanded daisy before i could reply that he was my neighbour dinner was announced wedging his tense arm imperatively under mine tom buchanan compelled me from the room as though he were moving a checker to another square slenderly languidly their hands set lightly on their hips the two young women preceded us out on to a rosycoloured porch open toward the sunset where four candles flickered on

GENERATED text: the garden and ella blinds and mashed potatoes and coffee i dipped mansion in the movingpicture pressure of his hand appeared with the pyramid of

--------------------------

Deep LSTM Model Text: 4:

SEED text: flock of white sheep turn the corner i said have to leave you you interposed tom quickly be hurt if you come up to the apartment you she urged telephone my sister catherine said to be very beautiful by people who ought to like to we went on cutting back again over the park toward the west hundreds at street the cab stopped at one slice in a long white cake of apartmenthouses throwing a regal homecoming glance around the neighbourhood mrs wilson gathered up her dog and her other purchases and went haughtily in going to have the mckees come

GENERATED text: and enormous affection and introductions crisp and shawls salons the shortcut i had been urging before at his feet mckees was startled baker invited him

--------------------------

Deep LSTM Model Text: 5:

SEED text: tell your chauffeur to go far away and spend an back in an hour then in a grave murmur name is the gasoline affect his think she said innocently we went in to my overwhelming surprise the livingroom was deserted i exclaimed she turned her head as there was a light dignified knocking at the front door i went out and opened it gatsby pale as death with his hands plunged like weights in his coat pockets was standing in a puddle of water glaring tragically into my eyes with his hands still in his coat pockets he stalked by me

GENERATED text: and selfish at the plane and shake the chapter of paternal wondered him eventually break to play in the betting creaking pivot and mineral had

--------------------------

Deep LSTM Model Text: 6:

SEED text: obscene word scrawled by some boy with a piece of brick stood out clearly in the moonlight and i erased it drawing my shoe raspingly along the stone then i wandered down to the beach and sprawled out on the sand most of the big shore places were closed now and there were hardly any lights except the shadowy moving glow of a ferryboat across the sound and as the moon rose higher the inessential houses began to melt away until gradually i became aware of the old island here that flowered once for dutch fresh green breast of the new

GENERATED text: town of burns metallic turbaned ones i had been urging before to play in tournaments and he proposed him of the weary bottle of enormous

--------------------------

Deep LSTM Model Text: 7:

SEED text: flashed at me indignantly turned around in the door and says let that waiter take away my then he went out on the sidewalk and they shot him three times in his full belly and drove of them were i said remembering with his nostrils turned to me in an interested way understand looking for a business the juxtaposition of these two remarks was startling gatsby answered for me he exclaimed the mr wolfshiem seemed disappointed is just a friend i told you talk about that some other beg your said mr wolfshiem had a wrong a succulent hash arrived and

GENERATED text: i was a tiny strawhaired and and wellconcealed was a solid sticky bob of oboes and southern tattle to the incessant pressure of the walls

--------------------------

Deep LSTM Model Text: 8:

SEED text: to kiss me any time during the evening nick just let me know and be glad to arrange it for you just mention my name or present a green card giving out suggested gatsby looking around having a must see the faces of many people heard arrogant eyes roamed the crowd go around very he said fact i was just thinking i know a soul you know that gatsby indicated a gorgeous scarcely human orchid of a woman who sat in state under a whiteplum tree tom and daisy stared with that peculiarly unreal feeling that accompanies the recognition of a

GENERATED text: century was the frogs doors of the fadedgilt brothel and saloon the drawbridge the younger of the bracing undeserted celebrity of the oculist set his

--------------------------

Deep LSTM Model Text: 9:

SEED text: a stable out of a wants to go to demanded daisy insistently eyes floated toward her she cried look so their eyes met and they stared together at each other alone in space with an effort she glanced down at the table always look so she repeated she had told him that she loved him and tom buchanan saw he was astounded his mouth opened a little and he looked at gatsby and then back at daisy as if he had just recognized her as someone he knew a long time ago resemble the advertisement of the she went on innocently

GENERATED text: tells him i was a few smallbreasted hand of violence of the towel named tragically down at the yard orders we was a oxford of

--------------------------

Deep LSTM Model Text: 10:

SEED text: believed he was lying he hurried the phrase at or swallowed it or choked on it as though it had bothered him before and with this doubt his whole statement fell to pieces and i wondered if there something a little sinister about him after all part of the middle i inquired casually family all died and i came into a good deal of his voice was solemn as if the memory of that sudden extinction of a clan still haunted him for a moment i suspected that he was pulling my leg but a glance at him convinced me otherwise

GENERATED text: and if i had desired him and he loves him i had rejoin it a introduction that merton the moment i noticed to abandon you

--------------------------

Model 2’s generated text seems to suffer from similar issues as the first one. It understands general grammar and what types of words go after what, however it isn’t really saying anything. This one also sounds a lot less like the author’s writing compared to the original model. It completely seems like the model didn’t have enough training time to fully grasp the language and context, which could have been solved with more epochs.

These aggregated results make it obvious to me that the original model outperformed the second more complex one. They both struggle to actually add information to the story, however the first one generates sentences that just make more sense and are more understandable. The second one doesn’t transition between any ideas at all, it just predicts almost it seems like a random word.

# **Reflection**

I’ve noticed that like always, data cleaning and preparation is the the most difficult part about implementing text generation models. I will most definitely be better prepared for projects in this realm of text generation in the future since this one took me a while to get set up. Understanding what was going on in the data preparation portion of the assignment was difficult.

For the future, I likely will include more nodes in each layer as well as run the model training for at least 100 epochs next time, likely closer to 250 to ensure that the complexity of the relationships the model can grasp are fully optimally trained.

Through working on this assignment, I also thought heavily about the applications of a model like this since it wasn’t really intuitive to me at the beginning. I could see an exceptionally well performing model like this being applied to help authors write sequels of their novels to assist in giving ideas and general sentence structure for the author to work with. I also see this applying to search engines or autocomplete on text messages.