It was in 2016 when [Andy Herd generated new *Friends* scenes](https://twitter.com/_Pandy/status/689209034143084547) by training a recurrent neural network on all the show’s episodes. Herd’s work went pretty viral at the time and I thought:

[via GIPHY](https://giphy.com/gifs/inpulsedm-3oEjHTazrdiQeYaR68)

And also:

[via GIPHY](https://giphy.com/gifs/CDoxe35inxhfO)

At the time I dabbled a bit with [Andrej Karpathy’s tutorials](http://karpathy.github.io/2015/05/21/rnn-effectiveness/) for character-level RNNs; his work and tutorials undergird a lot of the kind of STUNT TEXT GENERATION work we see in the world. Python is not my strongest language, though, and I did not ever have a real motivation to understand the math of what was going on. I watched the masters like [Janelle Shane](https://twitter.com/JanelleCShane) instead.

[TensorFlow for R](https://tensorflow.rstudio.com/) has changed that for me. Not only is the R interface that RStudio has developed just *beautiful*, but now these fun text generation projects provide a step into understanding how these neural networks model work at all, and deal with text in particular. Let’s step through how to take the text of *Pride and Prejudice* and generate new Jane-Austen-esque text.

This code borrows heavily from a couple of excellent sources.

* [Jonathan Nolis’ project on offensive license plates](https://github.com/jnolis/banned-license-plates) (That link is for their code; you can read a [great narrative explanation as well](https://towardsdatascience.com/using-deep-learning-to-generate-offensive-license-plates-619b163ed937).)
* [RStudio’s example code for text generation](https://keras.rstudio.com/articles/examples/lstm_text_generation.html)

Before starting, you will need to install keras so be sure to check out [details on installation](https://keras.rstudio.com/reference/install_keras.html).

**Tokenize**

We are going to train a character-level language model, which means the model will take a single character and then predict what character should come next, based on the ones that have come before. First step? We need to take *Pride and Prejudice* and divide it up into individual characters.

[via GIPHY](https://giphy.com/gifs/bbc-drama-pride-and-prejudice-l4JyY6IdblsismaeA)

The code below *keeps* both capital and lowercase letters, and builds a model that learns when to use which one. This is computationally more intensive than training a model that only learns about the letters themselves in lower case; if you want to start off with that kind of model, change to the default behavior for tokenize\_characters() of lowercase = TRUE.

library(keras)

library(tidyverse)

library(janeaustenr)

library(tokenizers)

max\_length <- 40

text <- austen\_books() %>%

filter(book == "Pride & Prejudice") %>%

pull(text) %>%

str\_c(collapse = " ") %>%

tokenize\_characters(lowercase = FALSE, strip\_non\_alphanum = FALSE, simplify = TRUE)

print(sprintf("Corpus length: %d", length(text)))

## [1] "Corpus length: 684767"

chars <- text %>%

unique() %>%

sort()

print(sprintf("Total characters: %d", length(chars)))

## [1] "Total characters: 74"

A good start!

**CHOP CHOP CHOP**

Next we want to cut the whole text into pieces: sequences of max\_length characters. These will be the chunks of text that we use for training.

dataset <- map(

seq(1, length(text) - max\_length - 1, by = 3),

~list(sentence = text[.x:(.x + max\_length - 1)],

next\_char = text[.x + max\_length])

)

dataset <- transpose(dataset)

**Vectorize**

Now it’s time to make a big set of vectors of these chunks of text. If you make max\_length larger, this vectors object can get unwieldy in terms of memory.

vectorize <- function(data, chars, max\_length){

x <- array(0, dim = c(length(data$sentence), max\_length, length(chars)))

y <- array(0, dim = c(length(data$sentence), length(chars)))

for(i in 1:length(data$sentence)){

x[i,,] <- sapply(chars, function(x){

as.integer(x == data$sentence[[i]])

})

y[i,] <- as.integer(chars == data$next\_char[[i]])

}

list(y = y,

x = x)

}

vectors <- vectorize(dataset, chars, max\_length)

**Model definition**

So far all we’ve been doing is chopping text into bits and rearranging data structures. Finally, it is time to delve into ❇ DEEP LEARNING ❇. The first step is to create a model. I’ve used the same parameters as [the RStudio LSTM example](https://keras.rstudio.com/articles/examples/lstm_text_generation.html); this next step is fast as it is only defining the kind of model architecture we are going to use.

create\_model <- function(chars, max\_length){

keras\_model\_sequential() %>%

layer\_lstm(128, input\_shape = c(max\_length, length(chars))) %>%

layer\_dense(length(chars)) %>%

layer\_activation("softmax") %>%

compile(

loss = "categorical\_crossentropy",

optimizer = optimizer\_rmsprop(lr = 0.01)

)

}

Let’s also make a function that fits the model for a set number of epochs.

fit\_model <- function(model, vectors, epochs = 1){

model %>% fit(

vectors$x, vectors$y,

batch\_size = 128,

epochs = epochs

)

NULL

}

**Model training & results**

Now it’s almost time to **train** the model on our data. Let’s make some more functions.

This one generates a phrase from a model, text, set of characters, and parameters like the maximum length of phrase and diversity, i.e. how WILD we are going to let the model be.

generate\_phrase <- function(model, text, chars, max\_length, diversity){

# this function chooses the next character for the phrase

choose\_next\_char <- function(preds, chars, temperature){

preds <- log(preds) / temperature

exp\_preds <- exp(preds)

preds <- exp\_preds / sum(exp(preds))

next\_index <- rmultinom(1, 1, preds) %>%

as.integer() %>%

which.max()

chars[next\_index]

}

# this function takes a sequence of characters and turns it into a numeric array for the model

convert\_sentence\_to\_data <- function(sentence, chars){

x <- sapply(chars, function(x){

as.integer(x == sentence)

})

array\_reshape(x, c(1, dim(x)))

}

# the inital sentence is from the text

start\_index <- sample(1:(length(text) - max\_length), size = 1)

sentence <- text[start\_index:(start\_index + max\_length - 1)]

generated <- ""

# while we still need characters for the phrase

for(i in 1:(max\_length \* 20)){

sentence\_data <- convert\_sentence\_to\_data(sentence, chars)

# get the predictions for each next character

preds <- predict(model, sentence\_data)

# choose the character

next\_char <- choose\_next\_char(preds, chars, diversity)

# add it to the text and continue

generated <- str\_c(generated, next\_char, collapse = "")

sentence <- c(sentence[-1], next\_char)

}

generated

}

Notice that we seed the first characters for the model to use for prediction with a real chunk of text from *Pride and Prejudice*.

This next function fits the model to the set of vectors, and then generates phrases from the current version of the model.

iterate\_model <- function(model, text, chars, max\_length,

diversity, vectors, iterations){

for(iteration in 1:iterations){

message(sprintf("iteration: %02d ---------------\n\n", iteration))

fit\_model(model, vectors)

for(diversity in c(0.2, 0.5, 1)){

message(sprintf("diversity: %f ---------------\n\n", diversity))

current\_phrase <- 1:10 %>%

map\_chr(function(x) generate\_phrase(model,

text,

chars,

max\_length,

diversity))

message(current\_phrase, sep="\n")

message("\n\n")

}

}

NULL

}

I’m sorry to say that we haven’t really done anything yet.

[via GIPHY](https://giphy.com/gifs/bbc-drama-pride-and-prejudice-26FLdSILnEZTBFPjO)

**Actually run the model**

But now! Now we are going to train the model.

How many times should you iterate through the model? You want to the loss to stabilize (lower is better) but once the loss is at whatever low value we can achieve for the data we have and the model architecture we have chosen, iterating more and more isn’t going to help anymore. For me with this data, about 40 iterations worked well.

model <- create\_model(chars, max\_length)

iterate\_model(model, text, chars, max\_length, diversity, vectors, 40)

## NULL

Now let’s see what we’ve got! Let’s look at several values for diversity, the measure for how creative/wacky we let the model be in which character to choose next in a sequence. We’ll try out values between 0.2 (less creative) and 0.6 (more creative).

result <- data\_frame(diversity = rep(c(0.2, 0.4, 0.6), 20)) %>%

mutate(phrase = map\_chr(diversity,

~ generate\_phrase(model, text, chars, max\_length, .x))) %>%

arrange(diversity)

result %>%

sample\_n(10) %>%

arrange(diversity) %>%

kable()

| **diversity** | **phrase** |
| --- | --- |
| 0.2 | sisters were the manner were so much the same family to be a moment to her sisters were not to see him and desired to her own party of the word, and the person was a man who had been a compliment of any of the subject. It was to be sure the subject of the satisfaction of the word, who cannot be well to be not betrod of the contrary to be sure the manner of the contrury on the part, and the sentiment of her sisters were so much and real the same proper and such a comfort of her sisters were at the day to be allow the family and such a word. If they were a disappointment of her family of the disappointment of the attention of the manner to her sisters were so much that I have not a few moment and a family to be all the manner of the consents of her family, who cannot do not been the manner t |
| 0.2 | s part, and was a contrive to her family to the discovery of the manner and the servance of her sisters were the person of the sentiment of the end of the subject of the manner to her family as they were all her attention of her family, and the manner of any of her sisters were of the family of the discovered to the evening to her aunt, and she could not be contrive with the subject, and she had been all the two distress of the contrury in the of her and the happiness of the sentence to her family of the contruct of her sisters were allow the consents of the evening the person were so after the connection of the contriculation of the discovered to be the subject of the attention of the consents and the servance of the subject of the sent the subject of the word, who had been a mind with th |
| 0.4 | fect a fortnight to be many of the connected to supplemer of an address to him in the subject. m\*va coseZ–and MZbZbZbZZZborness of the contrusion of her family, and was to supplecity of her sisters were needed to her attention with her behaviour of the minutes to much to prevent the minutes to her sisters were to determined to support to the proper an agitation of the manner and pain her sisters were easily were a follow a great ease to her attention of the proper any of the propers I am allow the same intention of the power to her aunt, who cannot all me to her delight to expect to the distress of the contricumed to controsted to see the persuaded her sisters were allow to please in a moment with the respect a money and contrusting her aunt to make him to be p |
| 0.4 | . Mr. Wickham had been a look and contrusment of her favourable attention was said, and in the family of the compliment of his great really in the perfect that they were of a money, and who had seen the place who was displeasure in the attentions had seen and heart of her lady in the after many aunt, and the means of the rest of the discheme to be allow to be a family, and she had soon afterwards paid it to have seen the distress of her friend and a disappointment of her connection to be not to be the sentiment of my mother was all the word, I have been a well had ever to such any as she had seen much of his sisters were heard and many and seeing her sisters were the different of the connections what you have determine in the subject. If they were of the as the consents of the evening of t |
| 0.4 | the happiness of her sisters and had heard of the earnest of the servance of the consequence of the family of the very little to her friend to her family of the lady of the moment to the persuaded her to her sisters were allow to the evening in the party to supple of the moment. It is to be sure the family of the sisters were allow the happy of him." "The evening and the of her contrary the manner and seeing the heart, and was so information of the appearance, the match he had not have a constant of the consents and seeing the dear the dear Elizabeth was exceedingly so to support to the attention of the distress of his favourable to be another to any of the happiness of the subject. The perfect to her own very little attention which he had been consequently at the fellowing her and absent |
| 0.4 | Mr. Bennet by Mr. Bennet to the hardsty frival of the will be a love the consents were of her sisters are to her aunt had heard and pleased to be any of her sisters were the morning of the happiness which made her and said, and she had not in the safisghing to the sent a considered to him as he had see her friend to her minutes with the son. I am conversation and heart of the others and the two man to be a great perfect it is much as the sisters became into the will she could not allow the evening all the attention of the subject of Mr. Wickham had been heard to her word of the minutes and manner she had been soon and his as I must be alamments which I have been a few moment she had been so contratient and conversation of her daughters were done in the evening with the persuaded the lady |
| 0.4 | pt the respect to be the intention of her own friend to be always to be allow the attentions was the attentions of the satisfaction of the principation in the happiness of the servance of the attention of the of her share to be allow the evening after the lady which the rest of the happiness of the return to make the seite and the discovered him with the after supers I thought I have seen Miss Bingley, and he should not inquiries of the having such a moment, and who could not be soon know me on the delighted to the consents of the next understant a moment, and was very little consents of her way with an each other she had been understand the party that he had not been a little of such an instantly and perfect the family and considered to make the moment of the lady to the hall, but when sh |
| 0.6 | de, and had the servant of the means and de Mr. Darcy was she said, they were parther, however, and pay to anxious and please any home to much unnost to him and wholly better of the real uncomfort and overtear to be this servance was seen and moment and as I am relating to her first and earnest of an instantly; and he had any person. If tuch, that he was met a sort of the moment. The word, defects to us the man for such a falling her sentiming to her carriage some periove they were we usual forward to the complete a condence of the country mention she was always cannot be allow to him had much the interest of money they she accomplice of prevent the mother to be so fondue to not be little she really one of her mother, and above all that he had heard of a different seeing the next was not t |
| 0.6 | he subject. I have good on him, to be as teach of the distress of Mr. Wickham she could not the evening after an stair. They she entered her sister as she had ever soon afterwards concerned to man as she had seemed to take to emong much well intention of the day of their various of her family; and at the nerselles which he is but she was the first to have some perfect and repressible of the mention of her aunt that I immank that you have been we was the satisfaction, and by Mr. Bingley, for the other; and you the pleasure in the appluce–not a great frecen in their case, indeed, and the contrirng to be said and attentive can was a moment he had no sertave in the hopes he entreat of me would be soon ressence, and was good manner that we cannot so the manner I am tclone, I think I am two muc |
| 0.6 | beth, she was seeing him to Mr. Darcy was the word she could have been her faurity of him, in the happiness in the of a word; and afterwards to take she had not eft to done in the common after only she had the connection was well, and she could not be for sent to be soon said in favowed to each other; that Mr. Bingley had only not complete of the speeched and to suppleter really as she should not hear he had been soon persuaded the consequently and repected that he had seen she had a mother with the dischements to lost the sort of the after a su;bee the discomposest and opened a pleasant considered to get of the poult to be ever seen and intention, and but the happy any moment to him the entroation of her sisters. She is in their sisters and some as a last who was really as he play to be s |

IT WAS TO BE SURE THE SUBJECT OF THE SATISFACTION OF THE WORD!

You can see here what the whole generated phrases look like, and notice how these are not complete sentences and would need some cleaning up from this state. If we’d like to pull out only complete sentences, we could do some text manipulation, sentence tokenization, etc.

**Conclusion**

Understanding how text generation works with deep learning and TensorFlow has been very helpful for me as I wrap my brain around these techniques more broadly. And that’s good, because exactly how practical of a skill is this, right?! I mean, [who needs to generate new text from an existing corpus in their day job???](https://stackoverflow.blog/2018/01/15/thanks-million-jon-skeet/)

Oh, that’s right: me. Ironically, when I did need to generate text in my day job, I turned to a [Markov chain generator](https://github.com/abresler/markovifyR). It is computationally less expensive and gives “nicer” results without lots of tuning; also I could guarantee that no user was going to be served any unintentionally offensive text. To sum up, if you have an immediate serioud need for text generation, I might recommend another method, but playing with text generation is a great way to understand deep learning. Let me know if you have any questions or feedback!