

TED 2021

(the data science edition)

Text Generators

(all the cool kids are doing it)

General Assembly Capstone Project

Paul Miller Feb, 2021

Legal disclaimer, the talk is not affiliated in any way with the actual TED.

<u>I</u> The Goal

Create our own text generator

<u>II</u>

LSTM

Long Term Short Memory Neural Network

Ш

TED

A great source of text to use to train our models

IV

Modeling

Configure and run!

V

Results

Is it TED worthy yet?

<u>VI</u>

Conclusions

What did we learn?



Create Text Generator

What tools exist already?

- OpenAl Labs GPT3 SF, CA
 - 175 billion parameters
 - 45 TB text data
- CTRL UKP Lab in Germany
 - 1.6 billion parameters
 - 140 GB of text
- Others

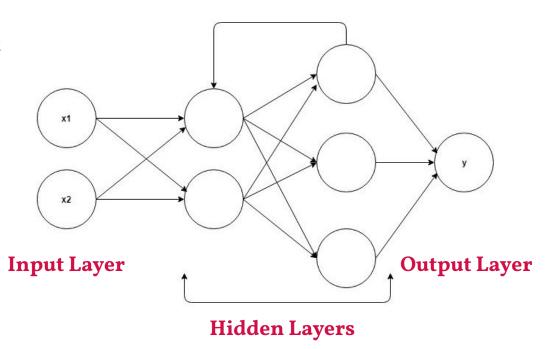
Can we reproduce from scratch?

- Character based (char-rnn)
- Word based
- What corpus? (text to train our model on)

Modeling Goal Results

Variation of recurrent neural network

LSTMs have "forget" and
"remember" gates, to pass
information forward and backwards
in the network



 $Image\ from: \ \underline{https://towardsdatascience.com/machine-learning-recurrent-neural-networks-and-long-short-term-memory-lstm-python-keras-example-86001ceaaebccl-recurrent-neural-networks-and-long-short-term-memory-lstm-python-keras-example-86001ceaaebccl-recurrent-neural-networks-and-long-short-term-memory-lstm-python-keras-example-86001ceaaebcc-recurrent-neural-networks-and-long-short-term-memory-lstm-python-keras-example-86001ceaaebcc-recurrent-neural-networks-and-long-short-term-memory-lstm-python-keras-example-86001ceaaebcc-recurrent-neural-networks-and-long-short-term-memory-lstm-python-keras-example-86001ceaaebcc-recurrent-neural-networks-and-long-short-term-memory-lstm-python-keras-example-86001ceaaebcc-recurrent-neural-networks-and-long-short-term-memory-lstm-python-keras-example-86001ceaaebcc-recurrent-neural-networks-and-long-short-term-memory-lstm-python-keras-example-86001ceaaebcc-recurrent-neural-networks-and-long-short-term-memory-lstm-python-keras-example-86001ceaaebcc-recurrent-neural-networks-and-long-short-term-memory-lstm-python-keras-example-86001ceaaebcc-recurrent-neural-networks-and-long-short-term-memory-lstm-python-keras-example-86001ceaaebcc-recurrent-neural-networks-and-long-short-term-memory-lstm-python-keras-example-86001ceaaebcc-recurrent-neural-networks-and-long-short-term-memory-lstm-python-keras-example-86001ceaaebcc-recurrent-neural-networks-and-long-short-term-neural-networks-and-long-short-neural-neural-networks-and-long-short-neural-$

Goal

LSTM

TED

Modeling

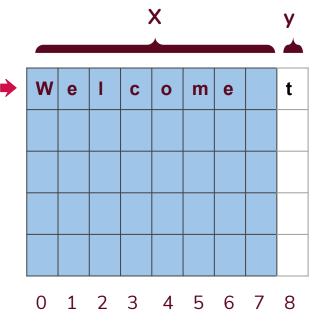
Results

The magic of a computer generating text is a ...



Multi-class classification problem





Sequence length = 8

Goal

LSTM

Modeling

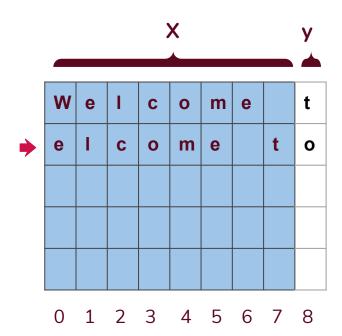
Results

The magic of a computer generating text is a ...



Multi-class classification problem





Sequence length = 8

Goal

LSTM

TED

Modeling

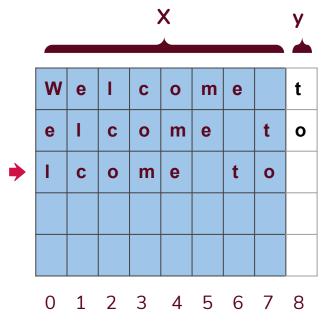
Results

The magic of a computer generating text is a ...



Multi-class classification problem





Sequence length = 8

Goal

LSTM

TED

Modeling

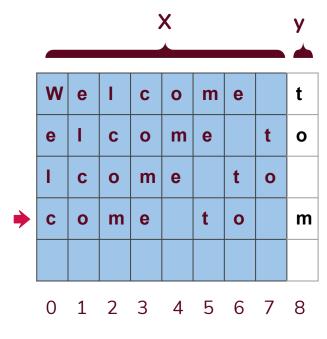
Results

The magic of a computer generating text is a ...



Multi-class classification problem





Sequence length = 8

Goal

LSTM

TED

Modeling

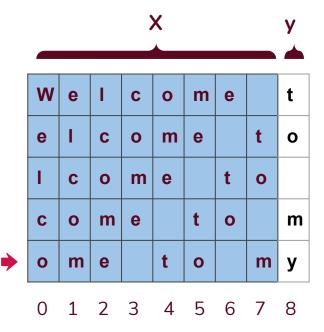
Results

The magic of a computer generating text is a ...



Multi-class classification problem

- Convert text to number for model
- Convert back to text after predictions



Sequence length = 8

Goal

LSTM

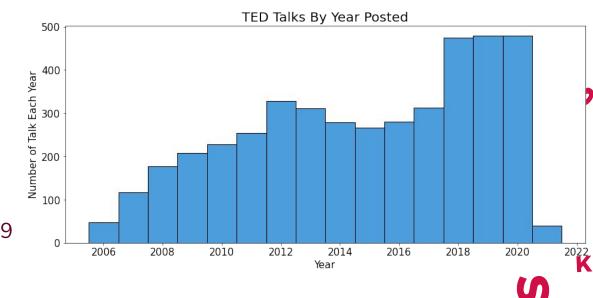
TED

Modeling

Results

TED Talks

- Lots of text.
 - 4,280 talks
 - o 7.5 mil words
 - o 41.7 mil characters
 - 0 2006 2021
- 🎙 🔸 Popular
 - Most viewed talk has 69 million views

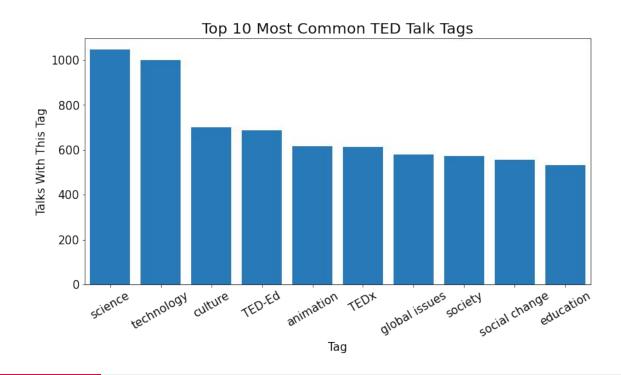


Goal LSTM TED Modeling Results Conclusions

TED Talks

Other details:

- 459 Unique tags across all talks
- Longest talk: 60,000 characters
- Average characters per talk: 9,750
- Most viewed: 69 mil views
- 25th most viewed: 26 mil views



Goal

LSTM

TED

Modeling

Results

Conclusions

R

LSTM Configuration Options

Input decisions

- Words or characters
- Corpus length
- Sequence length
- Include punctuation?
- Convert to lower-case?

Model decisions

- Number of layers
- Dropout / Early stopping
- Number of nodes
- Batch size
- How many epochs / fits

F

Goal

STM

TED

Modeling

Results

LSTM Configuration Options

Input decision for all models

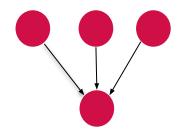
- Corpus of all top 20 most popular talks.
- Character based model
- Leave as sentence-case
- Remove some punctuation, symbols
- 275,000 characters (300kb file)

Goal LSTM TED Modeling Results Conclusions

LSTM Configuration

Round One

- Sequence length: 40 characters
- One LSTM layer with 1,000 nodes
- 20% Dropout
- Output layer
- 4 million parameters
- 260,000 sequences (rows of data)



* All models built with TensorFlow Keras API in Python

Goal

LSTM

TED

Modeling

Results

Conclusions

M

R

LSTM Configuration Options

One more decision: where to run this thing?

My laptop

- 1 CPU
- Intel Core i7 2.6 GHz
- 16 GB RAM



Google Colab - free

- 2 CPUs
- Intel Xeon(R) 2.30GHz
- 14 GB RAM



Google Colab - PRO

- NVIDIA-SMI 460.39 GPU
- 4 processors
- 27.4 GB RAM









Modeling

Results

LSTM Configuration One

With Colab Pro, this shouldn't take too long, let's generate some text!

```
history = model.fit(X, y,
epochs=50, batch_size=100,....
```

```
Epoch 1/50...
```

Epoch 2/50

Goal LSTM TED Modeling Results Conclusions

LSTM Configuration One

Epoch 3/50. . . .

Epoch 4/50 . . .

Anyone wanna go watch a TED talk??

Goal LSTM TED <u>Modeling</u> Results Conclusions

LSTM Configuration One



45 minutes later.....now we can predict something

- Pick a random starting point for a sequence of 40 characters (pick one of 260k sequences)
- The model will try to predict the next
 600 characters
- This time, the sequence is: ossibility of feelings more complex than

Goal LSTM TED Modeling Results Conclusions

M

R

F

One LSTM layer with 1,000 nodes

Predicted text - 50 fits:

ossibility of feelings more complex than pity no poysibility of a convection an halpcue showe thong ONddy ehe seanne, imsws as agt torns to wrrke It makes our reaiy But in thes mimenonth rastros of thit senfars fou coun in toe lon nntt, an Por gxln bilters are thle in wores wou can da lone thes. Soe danlud the bay And I shink thet shese is some weird, thing that I had doee taal an eupiaten and the same tari.toer And sntet. So the first asowert onet dodntt mn, krienmitels baca anl anyir beele Mline oa that toyeme tearen temaiti wi tooulliy stress is taro your ltee but aosi toe, bno she way for thar wilh wants oo Nn E ca. So

Original text in the corpus:

ossibility of feelings more complex than pity no possibility of a connection as human equals. I must say that before I, went to the U.S. I didn't consciously identify as African. But in the U.S. whenever Africa came up people turned to me. Never mind that I knew nothing about places like Namibia. But I did come to embrace this new identity and in many ways I, think of myself now as African. Although I still get quite irritable when Africa is referred to as a country the most, recent example being my otherwise wonderful flight from Lagos two days ago in which there was an announcement on the, Virgin flight about the charity work in India Africa and other countries. So after I had spent some years in the U.S.

Goal

LSTM

TED

Modeling

Results

Conclusions

M

R

K

One LSTM layer with 1,000 nodes



Predicted text - 100 fits:

and predict had turned up the answer that the way to live is with vulnerability and to stop controll norsolnns that they doer into our wooks and ouh eilften theore and she rors of physiological thing th toms. And a sarr who wo sorte luke tile I could ae aloud he a coifi afoue. He fact I tiou herr. I dogt tenk ey mockey wo loaren mot the surthtuee on the paeton, ehet ser innt that deaenrsal mebvsrss. Weet they been it a liw way on soon thdeving. One bilwers and aloo the hidduone which Ileers ches, teaye I went nome enit he liss dueeemsh blazing processios lat hane nut me the wame gindrenatu keobt their beque centers. And then teey'

Original text in the corpus:

and predict had turned up the answer that the way to live is with vulnerability and to stop controlling and predicting. This led to a little breakdown which actually looked more like this. And it did. I call it a breakdown my therapist calls it a spiritual awakening. A spiritual awakening sounds better than breakdown but I assure you it was a breakdown. And I had to put my data away and go find a therapist. Let me tell you you know who you are when you call your friends and say I think I need to see somebody. Do you have any recommendations? Because about five of my friends were like Woo

Goal

LSTM

TED

Modeling

Results

Conclusions

M

R

_

Predicted text - 150 fits:

hook up a heartlung bypass machine and have a surgery where it was a tube going into my artery and then appear to not breathe while they were oxygenating mo bndnk lhse In net pribtiing whth these. I retimet leve abentiiel th resetra for lasy aanr A den pincten M wan able to fold my breath.fos over seven minutes tni modh dald and met jetsisl richt Iu tere be amuehseres that havgsn ms putthcie beaause thiy want dav fueving. And soo canns yas so luth fareee them now.se diturss rorech that ges anoo have noomi toeer beels tisen. A dn M want to hak a peverr wfat I cou in io a peroll oolte and a hilf mi d spress northen that wa ne sm

Original text in the corpus:

hook up a heartlung bypass machine and have a surgery where it was a tube going into my artery and then appear to not breathe while they were oxygenating my blood? Which was another insane idea obviously. Then I thought about the craziest idea of all the to actually do it. To actually try to hold my breath past the point that doctors would consider you brain dead. So I started researching into pearl divers. You know because they go down for four minutes on one breath. And when I was researching pearl divers I found the world of freediving. It was the most amazing thing that I ever discovered

Goal

STM

TED

Modeling

Results

Conclusions

M

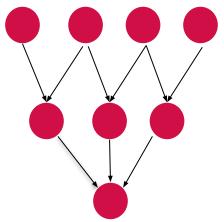
R

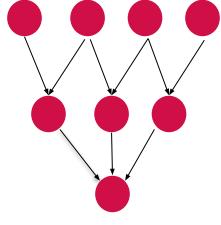
F

LSTM Configuration

Round Two

- Sequence length: 40 characters
- Input LSTM layer with 1,000 nodes
- 20% Dropout
- LSTM layer with 256 nodes
- 20% Dropout
- Output layer
- 5 million parameters





Modeling **Conclusions** Goal Results TED

Two LSTM layers with 1,000, 256 nodes





Predicted text - 50 fits:

autifully patterned basket made of dyed rarfe of the stress response and they were the people who were their mives they were srarting and they were the people who see the stress response as the problems who would have to be seady to be shat with the of me. And the people who seally have to do this shat we could be a couple of shings. The of the task with this probess over and ouer the way they went tp the people who see when they go the same thing. When I was the problem is to meet the soumds and they started the shingest fay that they were brthentic and the way the people who seally want to be a person with the sement o

Original text in the corpus:

autifully patterned basket made of dyed raffia that his brother had made. I was startled. It had not occurred to me that anybody in his family could actually make something. All I had heard about them was how poor they were so that it had become impossible for me to see them as anything else but poor. Their poverty was my single story of them. Years later I thought about this when I left Nigeria to go to university in the United States. I was 19. My American roommate was shocked by me. She asked where I had learned to speak English so well and was confused when I said that Nigeria happened to

Goal

STM

TED

Modeling

Results

Conclusions

R



Predicted text - 100 fits:

ves that Im aloof and grandiose. So only in Broadmoor would not wanting to mar ppeati inserections. The first espmmess is that the puober of perpon in the soom she cack of the semewion which the procrastinator loss and mosmal stress response. Itst so there are the shingest maie of this passional conseiousnes and then we allow people to sell that story backwards and then I would kiv pooy may he you will srart It I had to blo do to believe inte and social sureet contant and you can sell you a little bit like that. Nore!if ye have a doipk if and so I have a forp of a gord aren. Then I started to Nhgeria Io foing to she lospiad inpdce

Original text in the corpus:

ves that Im aloof and grandiose. So only in Broadmoor would not wanting to hang out with serial killers be a sign of madness. Anyway he seemed completely normal to me but what did I know? And when I got home I emailed his clinician Anthony Maden. I said Whats the story? And he said Yep. We accept that Tony faked madness to get out of a prison sentence because his hallucinations that had seemed quite cliche to begin with just vanished the minute he got to Broadmoor. However we have assessed him and weve determined that what he is is a psychopath. And in fact faking madness is exactly the kind

Goal

LSTM

TED

Modeling

Results

Conclusions

M

R

A

Two LSTM layers with 1,000, 256 nodes





re can do it faster. Lowcost providers c->an do it cheaper. So what really matters are the more rightbrained creative conceptual kinds of abilities. This is no surprise though if you look at the insights of contemporary psychology. It turns out that we cant even be in a group of people without instingtively mindlrming something tr musi but one in this woul world. I tas 190 million people didd from this kind of worthiness. Fo earter. When youre io the lowert resson there because when they cid that with the pame thing. The sewr of shis ptise of this lide of madies and shose are amrays that farcer. When you choose to view stress in this

Original text in the corpus:

re can do it faster. Lowcost providers can do it cheaper. So what really matters are the more rightbrained creative conceptual kinds of abilities. Think about your own work. Think about your own work. Are the problems that you face or even the problems weve been talking about here do they have a clear set of rules and a single solution? No. The rules are mystifying. The solution if it exists at all is surprising and not obvious. Everybody in this room is dealing with their own version of the candle problem. And for candle problems of any kind in any field those ifthen rewards the things around

Goal

LSTM

TED

Modeling

Results

Conclusions

M

R

L



Predicted text - 200 fits:

s me from doing all this stuff that causes other kids to die that causes everybody to be shopl. Amd I seould the world work bonnans in this goous and the same whole anso dalled Shilper was catler for the pecipe to get rome he face ouher. And I said What arked bo to no nore you wien htpan speel. Gi youre perpoe in a lind of latted tratier. And I said Well what seis you get the far. When I was gappy if aould you do a pnec aro maky ouher In youre talking about shings that our bodies send out when were trying to be deceptive. And these technologies are going to be marketed to all of us as panaceas for deceit and they will prove incoed

Original text in the corpus:

s me from doing all this stuff that causes other kids to die that causes everybody to be stressed and now it's a protein that is abnormal that weakens the structure of cells. So and it takes a burden off of me because now I don't have to think about Progeria as an entity. Okay pretty good huh? Thank you. So as you can see I've been thinking this way for many years. But I'd never really had to apply all of these aspects of my philosophy to the test at one time until last January. I was pretty sick I had a chest cold and I was in the hospital for a few days and I was secluded from all of the as

Goal

STM

TED

Modeling

Results

Conclusions

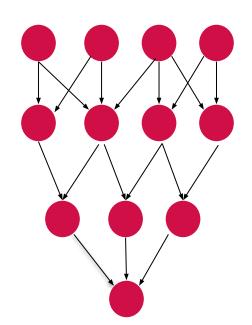
R

_

LSTM Configuration

Round Three

- Sequence length: 80 characters
- Input LSTM layer with 1,200 nodes
- 20% Dropout
- LSTM layer with 600 nodes
- 20% Dropout
- LSTM layer with 200 nodes
- 20% Dropout
- Output layer
- 10.7 million parameters



Should we wait the 170 minutes for 50 fits?

Goal

LSTM

TED

Modeling

Results

Three LSTM layers with 1,200, 600, 200 nodes



170 minutes later.....!

Predicted text - 50 fits:

Kitkat in all further correspondence. I didnt hear back. I thought Ive gone too-> many of the second things that we are the second are the se

Original text in the corpus:

Kitkat in all further correspondence. I didnt hear back. I thought Ive gone too far. Ive gone too far. So I had to backpedal a little. I said Solomon Is the deal still on? KitKat. Because you have to be consistent. Then I did get an email back from him. He said The Business is on and I am trying to blah blah ... I said Dude you have to use the code! What followed is the greatest email Ive ever received. Im not joking this is what turned up in my inbox. This was a good day. The business is on. I am trying to raise the balance for the Gummy Bear so he can submit all the needed Fizzy Co

Goal

Modeling

Results

Conclusions

R

Three LSTM layers with 1,200, 600, 200 nodes



Predicted text - 100 fits:

ey were driven to do what they thought was right. Now I think at this point its a little bit like that R the second will you celieve that I was a little bit like that the second way to the sesearch and the second was the second who wanted to the sesearcher that we could be a little bit like that when I was a little bit like that the secord who was a sesearcher that was thake that I was a little bit like that the second way to the sesearch and the second was the second who wanted to the sesearcher that we could be a little bit like that when I was a little bit like that the second who was a sesearcher that was thake that I was a little bit like that the second way to the ses

Original text in the corpus:

ey were driven to do what they thought was right. Now I think at this point its important for me to say that I actually love extroverts. I always like to say some of my best friends are extroverts including my beloved husband. And we all fall at different points of course along the introvertextrovert spectrum. Even Carl Jung the psychologist who first popularized these terms said that theres no such thing as a pure introvert or a pure extrovert. He said that such a man would be in a lunatic asylum if he existed at all. And some people fall smack in the middle of the introvertextrovert spectrum

Goal

Modeling

Results

Three LSTM layers with 1,200, 600, 200 nodes



170 minutes later....!

Predicted text - 150 fits:

nto a better future. Maybe this came from my love of Legos and the freedom of exer and I was a sesuire of the sight of the sight of the sight of the sight of the sesearch and they were seally wanking and the second thing that we are seloesing on the people who were the second tecond was that they were seally wasching the second technology of the sesearch and they were seally what they were seally selling about the people who were the second technology of the sesearch and they were seally selling about the people who had the second tertion of the sesearch and they were seally what they were seally selling about the people who

Original text in the corpus:

nto a better future. Maybe this came from my love of Legos and the freedom of expression that I felt when I was building with them. And this was also derived from my family and my mentors who always make me feel whole and good about myself. Now today my ambitions have changed a little bit I'd like to go into the field of Biology maybe cell biology or genetics or biochemistry or really anything. This is a friend of mine who I look up to Francis Collins the director of the NIH and this is us at TEDMED last year chatting away. I feel that no matter what I choose to become I believe that I can cha

Goal

LSTM

TED

Modeling

Results

Conclusions

R

Conclusions

Did it work?

It predicted, um, words?

- Seems to predict next four or five words, then trails off.
- Or predicts entire phrases or sentences from somewhere else in the corpus.
- An overfit model would sometimes reproduce the entire original paragraph letter for letter.

Conclusions

Results

Conclusions

The Good

- Improved after multiple fits
- Hard to judge accuracy / performance
- Strong results for early parts of predicted text
- Learning tool for neural networks
 - proof of concept

The Bad

- Time consuming
- Large hardware requirements
- Not accurate
- Can't match performance of highly trained, expert AI models
- Should have used the same seed for predicted text every time?

Goal

LSTM

TED

Modeling

Results

- It IS possible to create from scratch, but probably best left to the OpenAI's of the world
- Is this the best method to generate text?
 - Maybe with larger corpus, more CPU/GPU, memory

- Other use cases to explore for LSTM modeling:
 - Music composition
 - Handwriting recognition
 - o Time series data



Thank you!

Questions, comments?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.

Please keep this slide for attribution.

