News Generator

Swee Loke, Dec 2010
University of Toronto SCS 3546 Deep Learning

Original Inspiration

Want to predict user purchase behavior

Translate user orders into sentences, and each word represents an item and use text generator to predict the next item of interest….

Example

```
33120 28985 9327 45918 30035 17794 40141 1819 43668
1 33120 Organic Egg Whites
2 28985 Michigan Organic Kale
3 9327 Garlic Powder
4 45918 Coconut Butter
5 30035 Natural Sweetener
6 17794 Carrots
7 40141 Original Unflavored Gelatine Mix
8 1819 All Natural No Stir Creamy Almond Butter
9 43668 Classic Blend Cole Slaw


33754 24838 17704 21903 17668 46667 17461 32665
1 33754 Total 2% with Strawberry Lowfat Greek Strained Yogurt
2 24838 Unsweetened Almondmilk
3 17704 Lemons
4 21903 Organic Baby Spinach
5 17668 Unsweetened Chocolate Almond Breeze Almond Milk
6 46667 Organic Ginger Root
7 17461 Air Chilled Organic Boneless Skinless Chicken Breasts
8 32665 Organic Ezekiel 49 Bread Cinnamon Raisin
```

Lack of suitable datasets

- Tried instacart dataset - orders (3,346,083) and products (49,685)
- Products very specific, not generalized

400 products end with “milk”

product_id	product_name	aisle_id	department_id
329	Organic Whole Grassmilk Milk	84	16
432	Vanilla Almond Breeze Almond Milk	91	16
871	Rose & Apricot Hair Milk	22	11
877	Ultra-Filtered Whole Milk	84	16
1423	100% Lactose Free Whole Calcium Enriched Milk	91	16
...	...	...	...
49112	Lowfat Vanilla Milk	84	16
49319	Mozzarella String Cheese Made with 2% Reduced ...	21	16
49412	Original Cashew Milk	84	16
49517	0% Fat Free Organic Milk	84	16
49610	100% Lactose Free Fat Free Milk	91	16

330 rows × 3 columns

Order size varies

	max_items
count	3346083.000000
mean	10.107073
std	7.542326
min	1.000000
25%	5.000000
50%	8.000000
75%	14.000000
max	145.000000

Then decided to build a generic text generation model

- Can be trained on any text sequence
- Can support large datasets

Use a numbers of experiments to decide final model structure

- Comparing 2 datasets (5.8M characters each)
- Comparing max_sequence_length (10, 20, 30)
- Comparing using or not using pre-trained embeddings
- Comparing different pretrained embeddings
- Comparing different number of LSTM layer
- Comparing different RNN units per LSTM layer

Dataset 1: Short news

Just a headline and a short description

	headline	short_description
0	There Were 2 Mass Shootings In Texas Last Week, But Only 1 On TV	She left her husband. He killed their children. Just another day in America.
1	Will Smith Joins Diplo And Nicky Jam For The 2018 World Cup's Official Song	Of course it has a song.
2	Hugh Grant Marries For The First Time At Age 57	The actor and his longtime girlfriend Anna Eberstein tied the knot in a civil ceremony.
3	Jim Carrey Blasts 'Castrato' Adam Schiff And Democrats In New Artwork	The actor gives Dems an ass-kicking for not fighting hard enough against Donald Trump.
4	Julianna Margulies Uses Donald Trump Poop Bags To Pick Up After Her Dog	The "Dietland" actress said using the bags is a "really cathartic, therapeutic moment."

```
''there were 2 mass shootings in texas last week but only 1 on tv she left her husband he killed their children just another day in america will smith joins [UNK] and nic
ky jam for the 2018 world cups official song of course it has a song hugh grant marries for the first time at age 57 the actor and his longtime girlfriend anna [UNK] tied
the knot in a civil ceremony jim carrey blasts [UNK] adam schiff and democrats in new artwork the actor gives dems an [UNK] for not fighting hard enough against donald tr
ump [UNK] [UNK] uses donald trump poop bags to pick up after her dog the [UNK] actress said using the bags is a really [UNK] therapeutic moment morgan freeman devastated
that sexual harassment claims could undermine legacy it is not right to [UNK] horrific incidents of sexual assault with [UNK] compliments or humor he said in a statement
donald trump is lovin new mcdonalds jingle in tonight show bit its catchy all right what to watch on amazon prime thats new this week theres a great m… '
```

Dataset 2: Long news

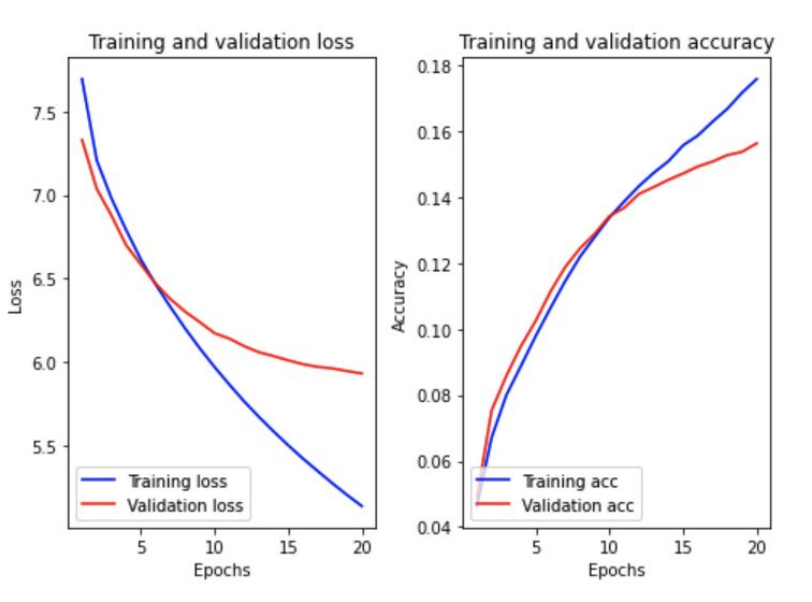
A headline and a text

	title	text
0	As U.S. budget fight looms, Republicans flip their fiscal script	The head of a conservative Republican faction in the U.S. Congress, who voted this month for a huge expansion of the national debt to pay for tax cuts, called himself a “fiscal conservative” on Sunday and urged budget restraint in 2018. In keeping with a sharp pivot under way among Republicans, U.S. Representative Mark Meadows, speaking on CBS’ “Face the Nation,” drew a hard line on federal s...
1	U.S. military to accept transgender recruits on Monday: Pentagon	Transgender people will be allowed for the first time to enlist in the U.S. military starting on Monday as ordered by federal courts, the Pentagon said on Friday, after President Donald Trump’s administration decided not to appeal rulings that blocked his transgender ban. Two federal appeals courts, one in Washington and one in Virginia, last week rejected the administration’s request to put ...
2	Senior U.S. Republican senator: 'Let Mr. Mueller do his job'	The special counsel investigation of links between Russia and President Trump’s 2016 election campaign should continue without interference in 2018, despite calls from some Trump administration allies and Republican lawmakers to shut it down, a prominent Republican senator said on Sunday. Lindsey Graham, who serves on the Senate armed forces and judiciary committees, said Department of Justic...
3	FBI Russia probe helped by Australian diplomat tip-off: NYT	Trump campaign adviser George Papadopoulos told an Australian diplomat in May 2016 that Russia had political dirt on Democratic presidential candidate Hillary Clinton, the New York Times reported on Saturday. The conversation between Papadopoulos and the diplomat, Alexander Downer, in London was a driving factor behind the FBI’s decision to open a counter-intelligence investigation of Moscow’...
4	Trump wants Postal Service to charge 'much more' for Amazon shipments	President Donald Trump called on the U.S. Postal Service on Friday to charge “much more” to ship packages for Amazon (AMZN.O), picking another fight with an online retail giant he has criticized in the past. “Why is the United States Post Office, which is losing many billions of dollars a year, while charging Amazon and others so little to deliver their packages, making Amazon richer and ...

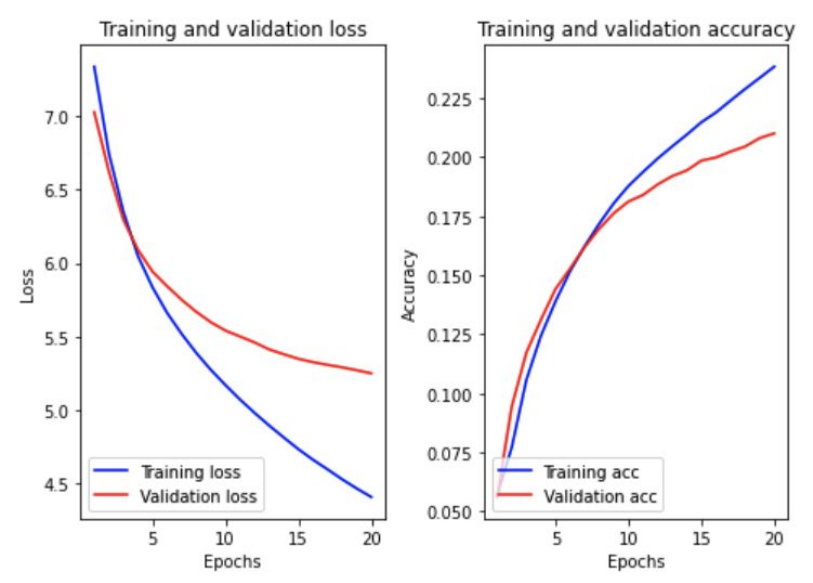
```
''as us budget fight looms republicans flip their fiscal script the head of a conservative republican faction in the us congress who voted this month for a huge expansion
of the national debt to pay for tax cuts called himself a fiscal conservative on sunday and urged budget restraint in 2018 in keeping with a sharp pivot under way among r
epublicans us representative mark meadows speaking on cbs face the nation drew a hard line on federal spending which lawmakers are bracing to do battle over in january wh
en they return from the holidays on wednesday lawmakers will begin trying to pass a federal budget in a fight likely to be linked to other issues such as immigration poli
cy even as the november congressional election campaigns approach in which republicans will seek to keep control of congress president donald trump and his republicans wa
nt a big budget increase in military spending while democrats also want proportional increases for nondefense discretionary spending on programs that … '
```

Datasets Comparison:

Short News:



Long News:



Comparing datasets

Using same text length (max_seq_len=10)

Short News:

Epoch 20/20

533/533 [==============================]
- 37s 69ms/step - loss: 5.2075 - accuracy:
0.1712 - val_loss: 5.9672 - val_accuracy: **0.1539**

Long News:

Epoch 20/20

527/527 [==============================]
- 37s 71ms/step - loss: 4.3932 - accuracy:
0.2397 - val_loss: 5.2444 - val_accuracy: **0.2106**

Using Long News for subsequent experiments

Generating Datasets

- Implemented classes: VocabClass, PretrainEmbedding and TextDataset
- One call should create TF datasets with train/test split

```
VALIDATION_SPLIT = 0.2
BATCH_SIZE = 128
EMBEDDING_DIM = 50
MAX_TOKENS = 20000 # this is fixed to 20K for all tests
TEXT_LIMIT = 5800000 # Using this as limit as it is just about the size of the short news text (so we will have both datasets about the same text length)

short_news_file_path = DATA_DIR+'short_news_text.txt'
short_news_dataset_10seq = TextDataset(short_news_file_path, TEXT_LIMIT, MAX_TOKENS, EMBEDDING_DIM, 10, VALIDATION_SPLIT, BATCH_SIZE)
```

```
Found 190189 word vectors.
Converted 18416 words (1584 misses) in the embedding

dataset size: 85282
num_validation_samples: 17056
dataset_train size: 68226
dataset_val size: 17056
dataset_train batch size: 533
dataset_val batch size: 133
```

Running experiment
- Simplify the different runs

```
[ ] # Running the test on long news dataset
    long_news_experiment_10seq = Experiment(
                        "str len 10 no pretrained embedding",
                        my_training_param,
                        EMBEDDING_DIM,
                        None, # if we don't pass in any, will use the untrained embedding (that is fully trainable)
                        True, # if we pass in embedding_matrix, do we want it to be trainable
                        long_news_dataset_10seq)

    EPOCH = 20
    long_news_result_10seq  = long_news_experiment_10seq.run_experiment(
                              f"Test first {EPOCH} epoch",
                              None,  # if we want to continue training from previous run, pass in the model
                              EPOCH )
```

```
dataset_train batch size: 527
dataset_val batch size: 131
Running str len 10 no pretrained embedding:Test first 20 epoch...

checkpoint_dir: ./training_checkpoints_12-01-03:24
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
empty-embedding (Embedding)  (128, None, 50)           1000100
_________________________________________________________________
lstm_2 (LSTM)                (128, None, 256)          314368
_________________________________________________________________
dropout_2 (Dropout)          (128, None, 256)          0
_________________________________________________________________
dense_2 (Dense)              (128, None, 20000)        5140000
=================================================================
Total params: 6,454,468
Trainable params: 6,454,468
Non-trainable params: 0
```

Max_seq_len: 10, 20 or 30

Using the same number of Epoch, max_seq_len=10 has better performance.

Suspect because it resulted in more datasets

Also less chance of non related text placed in 1 sequence.

```
<      News1       > <          News2        >

|  Seq 1 | seq 2 | seq 3 | seq 4 | …
```

Using Pre-trained Embeddings
(GloVe: Global Vectors for Word Representation)

- Compare using vs not using pre-trained embeddings
- Compare using 50dim, 200dim vs 300dim of GloVe embedding

300 dim embedding performs better

Comparing LSTM layers and RNN units

- Using 2 LSTM layers performs better
- Using more RNN units (1024 vs 512 vs 128) works better.

For the final model, we will use 1024 RNN units with 2 LSTM layers

Final Model

RNN unit = 1024

LSTM layer = 2

DROPOUT = 0.4

```
RNN_UNITS = 1024
DROPOUT = 0.4 # use higher drop out if needed more regularization
LSTM_LAYER = 2

my_training_param = {'rnn_units' : RNN_UNITS,
                     'dropout'   : DROPOUT,
                     'lstm_layer': LSTM_LAYER}

final_model = Experiment(
                  f"{LSTM_LAYER}LSTM dropout={DROPOUT}",
                  my_training_param,
                  EMBEDDING_DIM,
                  my_dataset.embedding.embedding_matrix, # if we don't pass in any, will use the untrained embedding (that is fully trainable)
                  False, # Not retrainable
                  my_dataset)

EPOCH = 100
result1  = final_model.run_experiment(
                        f"Test first {EPOCH} epoch",
                        None,  # if we want to continue training from previous run, pass in previous result dictionary
                        EPOCH )
```

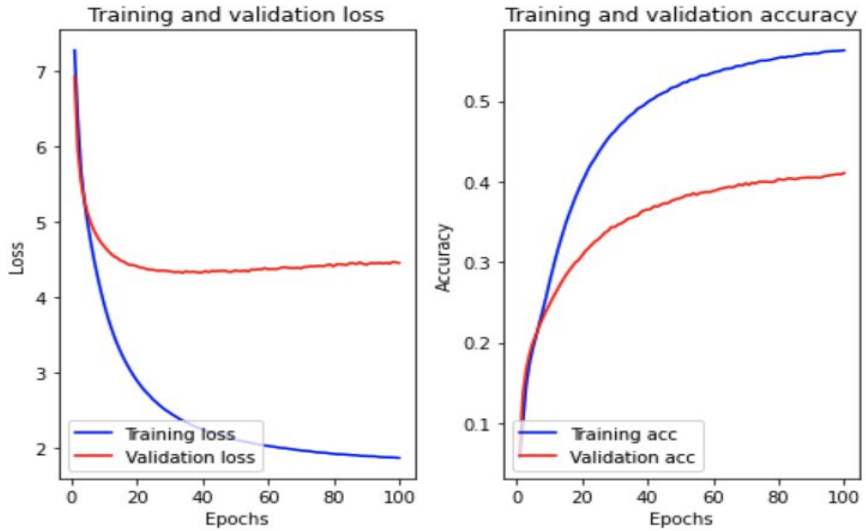
```
dataset_train batch size: 527
dataset_val batch size: 131
Running 2LSTM dropout=0.4:Test first 100 epoch...

checkpoint_dir: ./training_checkpoints_12-01-17:51
(20002, 300)
Model: "sequential_36"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
pretrained-embedding (Embedd (128, None, 300)          6000600
_________________________________________________________________
lstm_58 (LSTM)               (128, None, 1024)         5427200
_________________________________________________________________
dropout_58 (Dropout)         (128, None, 1024)         0
_________________________________________________________________
lstm_59 (LSTM)               (128, None, 1024)         8392704
_________________________________________________________________
dropout_59 (Dropout)         (128, None, 1024)         0
_________________________________________________________________
dense_34 (Dense)             (128, None, 20000)        20500000
=================================================================
Total params: 40,320,504
Trainable params: 34,319,904
Non-trainable params: 6,000,600
```

Model Performance

Metric = accuracy of exact match

```
Experiment: 2LSTM dropout=0.4:Test first 100 epoch
training_param: {'rnn_units': 1024, 'dropout': 0.4, 'lstm_layer': 2}
max_seq_len: 10
```




```
Epoch 95/100
527/527 [==============================] - 32s
61ms/step - loss: 1.8791 - accuracy: 0.5613 -
val loss: 4.4485 - val_accuracy: 0.4073
Epoch 96/100
527/527 [==============================] - 32s
61ms/step - loss: 1.8767 - accuracy: 0.5617 -
val loss: 4.4585 - val_accuracy: 0.4080
Epoch 97/100
527/527 [==============================] - 33s
63ms/step - loss: 1.8765 - accuracy: 0.5620 -
val loss: 4.4471 - val_accuracy: 0.4085
Epoch 98/100
527/527 [==============================] - 32s
61ms/step - loss: 1.8756 - accuracy: 0.5620 -
val loss: 4.4678 - val_accuracy: 0.4091
Epoch 99/100
527/527 [==============================] - 33s
63ms/step - loss: 1.8725 - accuracy: 0.5625 -
val loss: 4.4661 - val_accuracy: 0.4088
Epoch 100/100
527/527 [==============================] - 32s
61ms/step - loss: 1.8689 - accuracy: 0.5629 -
val_loss: 4.4558 - val_accuracy: 0.4106
```

Generated news with start strings:

```
start_string_list = ['Today is a good',
                     'Trump is',
                     'US plans to',
                     'Small business suffers',
                     'Tomorrow we will know',
                     'there were 2 mass shootings in texas last week',
                     'US stock market will continue to',
                     'The Oscar goes to',
                     "What is",
                     "Gunmen Assassinate Iran’s Top Nuclear Scientist",
                     "Iran’s president blames Israel for killing nuclear scientist and vows to",
                     "Man linked the killing of",
                     "This Thanksgiving"]
```

...and the results...

```
#1:  Today is a good...
        deal us technology with its new hampshires secretary jim mattis on wednesday clearing a relaxation new emergency management prevezons oversight
#2:  Trump is...
        expected with homeland security issues such a lucrative business income tax legislation that he stirred around the president donald trump
#3:  US plans to...
        block grants only the stunt that the younger older offpatent 63 percent of the senate if they could not a
#4:  Small business suffers...
        that was forced to a warrant before the notorious attacks and whitefishs deal and others it offering than half 800000
#5:  Tomorrow we will know...
        what at the resolution to deal and militias against him there has currently in 2013 said the senators on a
#6:  there were 2 mass shootings in texas last week...
        asked the weapons and it for several congressional aides say it us territories on us senate confirms charges that exist
#7:  US stock market will continue to...
        be party divisions a party computers with the details if flynn had opened fire hindering jobcreating said lawmakers have fled
#8:  The Oscar goes to...
        be able to see their users voiced misgivings misgivings about president signed by the united states until on slashing red
#9:  What is...
        known as the countrys presidential election the south carolina oregon and largest forecast an agreement i actually qualified to cuts
#10:  Gunmen Assassinate Iran's Top Nuclear Scientist...
        said mark zuckerberg to protect consumers to stay with the discussions with the indictment the 11 are into the united
#11:  Iran's president blames Israel for killing nuclear scientist and vows to...
        spur the latest in its old guidance in addition provision was passed by russia and patty murray of the mtcr
#12:  Man linked the killing of...
        more than an openended regulations and means committee ioc in july 3 from the treasury did not necessary rights values
#13:  This Thanksgiving...
        trump told the referee and companies said they believe even though they had concerns appropriately to working with the president
```

Try again..

```
#1:  Today is a good...
        deal were not immediately respond to block an industry officials
#2:  Trump is...
        expected in the federal law requires 60 percent of supreme
#3:  US plans to...
        cut the last week although adding for tax rate for
#4:  Small business suffers...
        that owns directv on its bills a housesenate senate that
#5:  Tomorrow we will know...
        what were receiving support conservative republicans and tweeted and white
#6:  there were 2 mass shootings in texas last week...
        abbott on wednesday to repeated meeting with a dinner room
#7:  US stock market will continue to...
        be hardpressed and republican effort to express condolences satisfaction in
#8:  The Oscar goes to...
        wall the company att incs program that would destroy the
#9:  What is...
        for engaging as the world for childhood arrivals daca mansion
#10:  Gunmen Assassinate Iran’s Top Nuclear Scientist...
        in us president on solid different tack on thursday the
#11:  Iran’s president blames Israel for killing nuclear scientist and vows to...
        destroy north koreas biggest impact earlier this weekend puts any
#12:  Man linked the killing of...
        the state rex tillerson from deportation of tribune trump would
#13:  This Thanksgiving...
        as the governments needs to cooperate with north korean leader
```

Conclusion

- The generated text is somewhat coherent, not completely random.

  (although not making complete sense either)

- More regularization is needed so we can improve the validation accuracy
- The model can be further enhance when training with different datasets

Thanks!

Github: https://github.com/sweeloke/news_generator