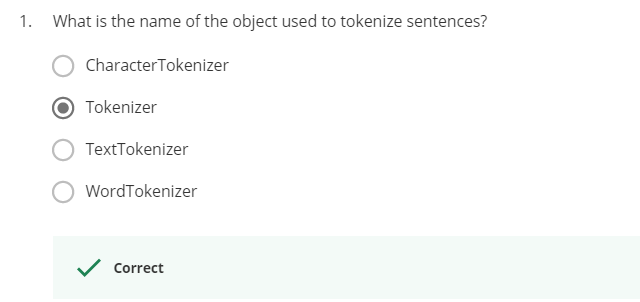
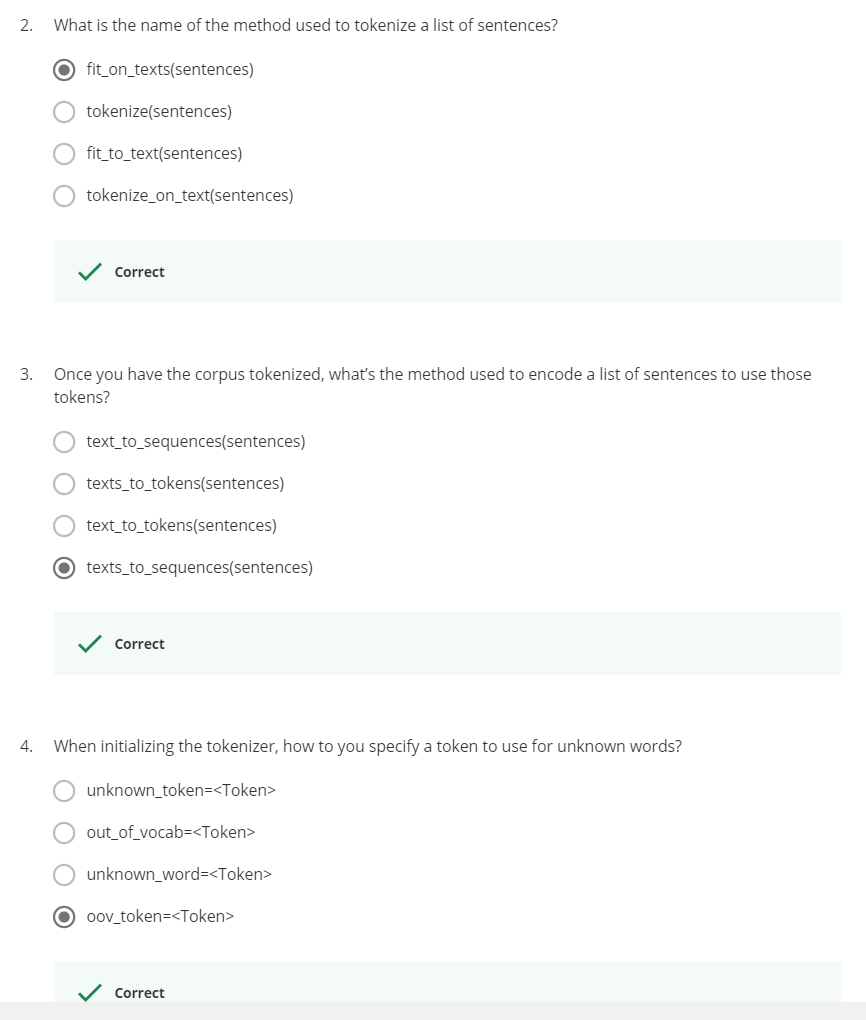
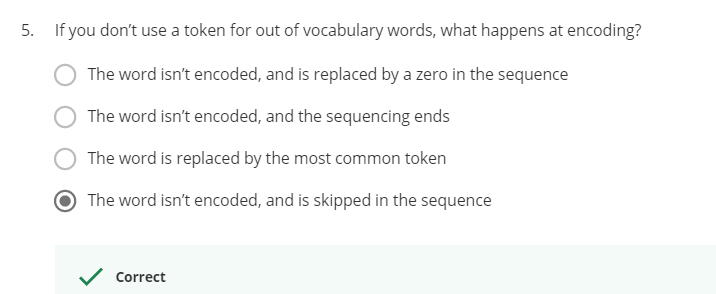
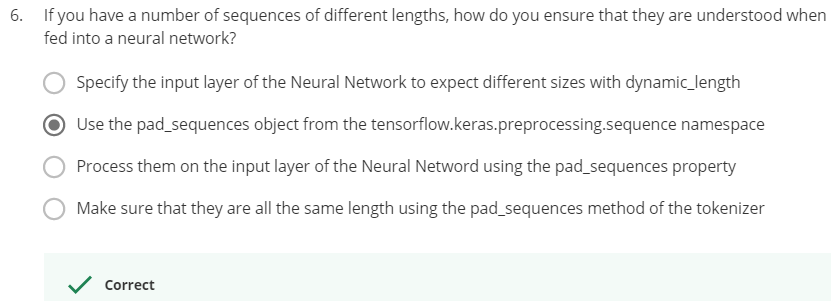
NLP in Tensorflow Study Notes

By Jane Huang









import tensorflow as tf

from tensorflow import keras

from tensorflow.keras.preprocessing.text import Tokenizer

from tensorflow.keras.preprocessing.sequence import pad\_sequences

sentences = [

    'I love my dog',

    'I love my cat',

    'You love my dog!',

    'Do you think my dog is amazing?'

]

tokenizer = Tokenizer(num\_words = 100, oov\_token="<OOV>")

tokenizer.fit\_on\_texts(sentences)

word\_index = tokenizer.word\_index

sequences = tokenizer.texts\_to\_sequences(sentences)

padded = pad\_sequences(sequences, maxlen=5)

print("\nWord Index = " , word\_index)

print("\nSequences = " , sequences)

print("\nPadded Sequences:")

print(padded)

# Try with words that the tokenizer wasn't fit to

test\_data = [

    'i really love my dog',

    'my dog loves my manatee'

]

test\_seq = tokenizer.texts\_to\_sequences(test\_data)

print("\nTest Sequence = ", test\_seq)

padded = pad\_sequences(test\_seq, maxlen=10)

print("\nPadded Test Sequence: ")

print(padded)

###############################################################

Word Index = {'<OOV>': 1, 'my': 2, 'love': 3, 'dog': 4, 'i': 5, 'you': 6, 'cat': 7, 'do': 8, 'think': 9, 'is': 10, 'amazing': 11}

Sequences = [[5, 3, 2, 4], [5, 3, 2, 7], [6, 3, 2, 4], [8, 6, 9, 2, 4, 10, 11]]

Padded Sequences:

[[ 0 5 3 2 4]

[ 0 5 3 2 7]

[ 0 6 3 2 4]

[ 9 2 4 10 11]]

Test Sequence = [[5, 1, 3, 2, 4], [2, 4, 1, 2, 1]]

Padded Test Sequence:

[[0 0 0 0 0 5 1 3 2 4]

[0 0 0 0 0 2 4 1 2 1]]

###############################################################

!wget --no-check-certificate \

    https://storage.googleapis.com/laurencemoroney-blog.appspot.com/sarcasm.json \

    -O /tmp/sarcasm.json

import json

with open("/tmp/sarcasm.json", 'r') as f:

    datastore = json.load(f)

sentences = []

labels = []

urls = []

for item in datastore:

    sentences.append(item['headline'])

    labels.append(item['is\_sarcastic'])

    urls.append(item['article\_link'])

from tensorflow.keras.preprocessing.text import Tokenizer

from tensorflow.keras.preprocessing.sequence import pad\_sequences

tokenizer = Tokenizer(oov\_token="<OOV>")

tokenizer.fit\_on\_texts(sentences)

word\_index = tokenizer.word\_index

print(len(word\_index))

print(word\_index)

sequences = tokenizer.texts\_to\_sequences(sentences)

padded = pad\_sequences(sequences, padding='post')

print(padded[0])

print(padded.shape)

###############################################################

[ 308 15115 679 3337 2298 48 382 2576 15116 6 2577 8434

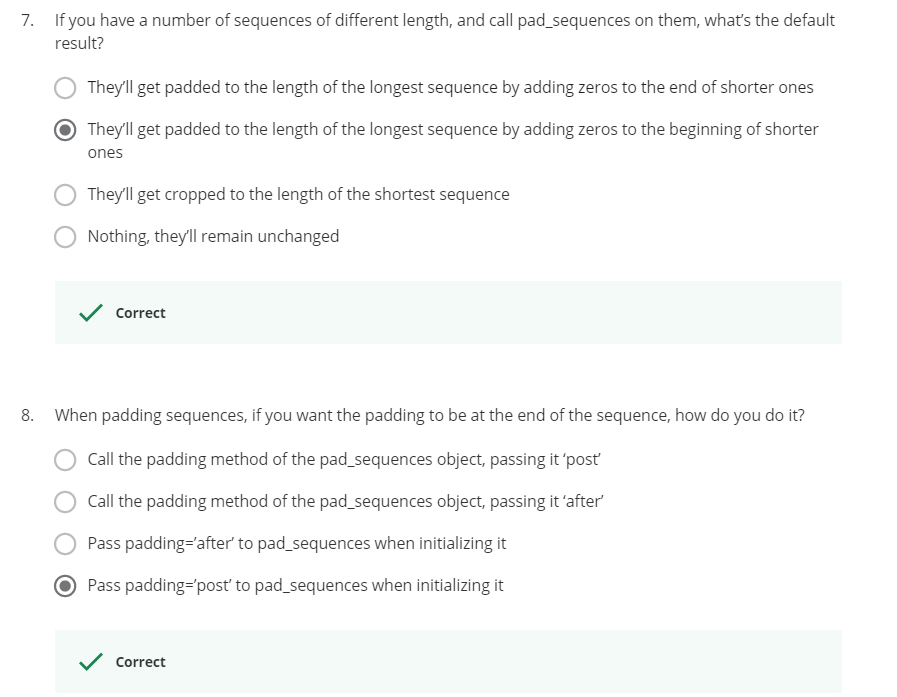
0 0 0 0 0 0 0 0 0 0 0 0

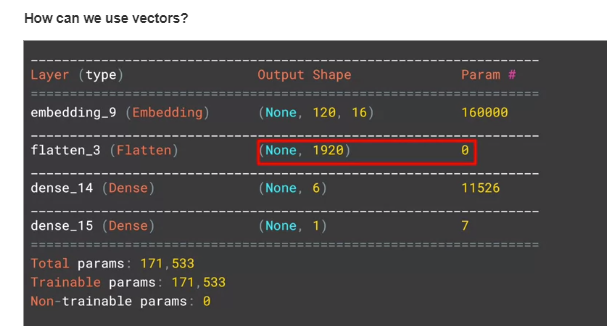
0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0]

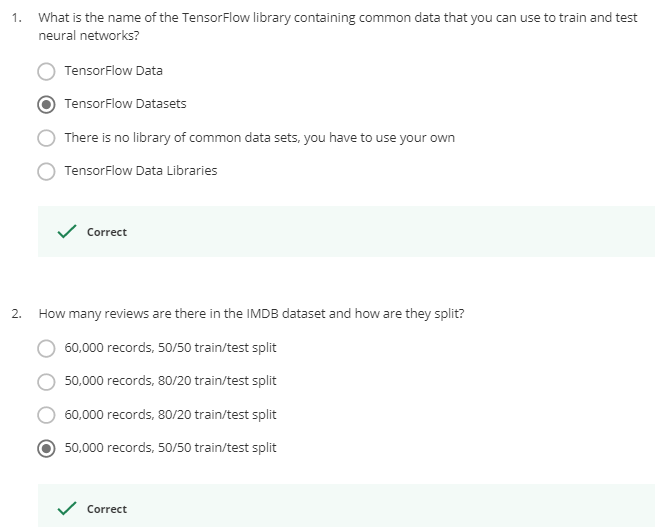
(26709, 40)

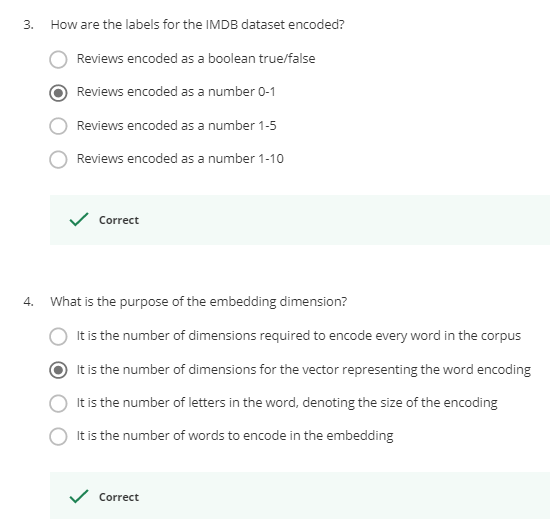
###############################################################

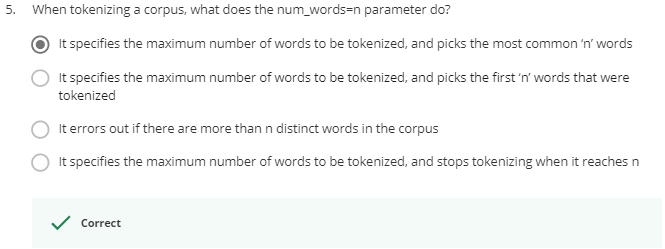


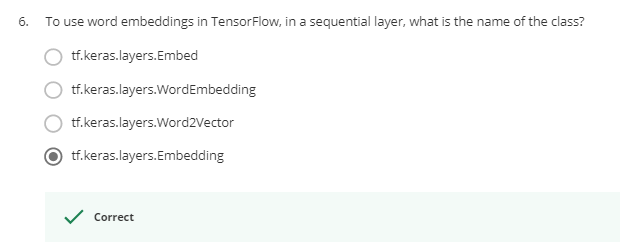


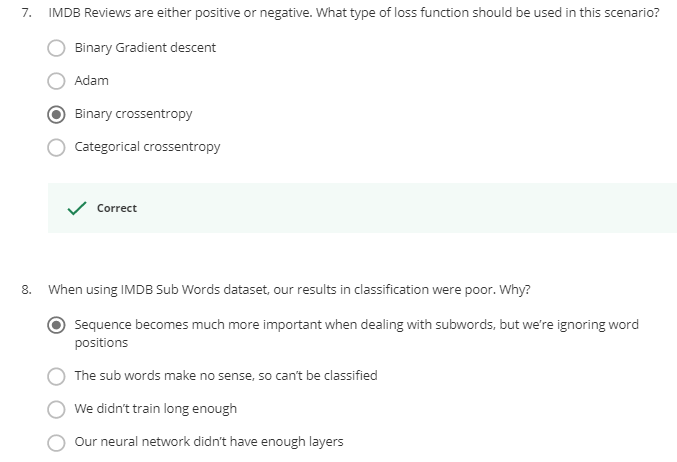
Week 2



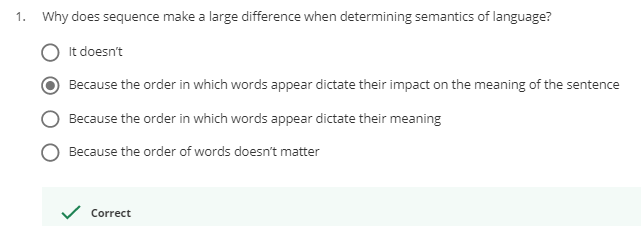


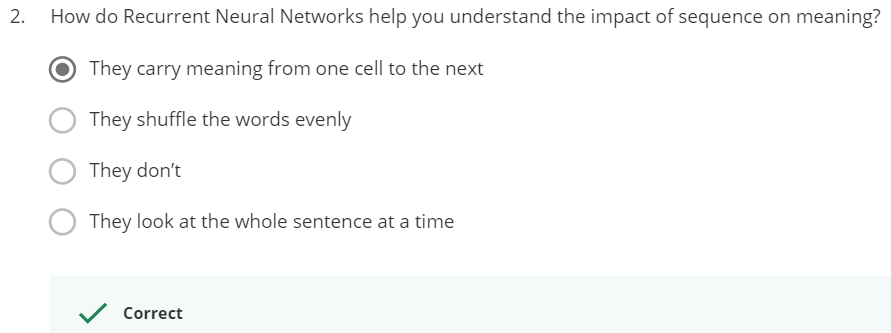


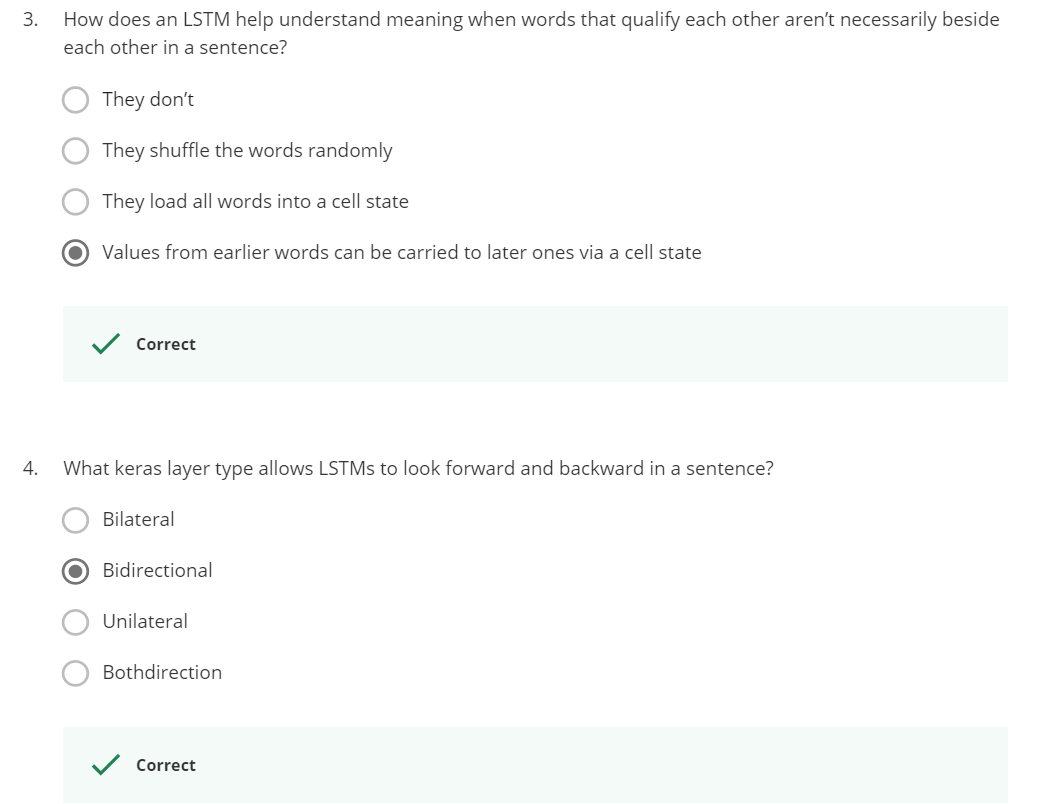


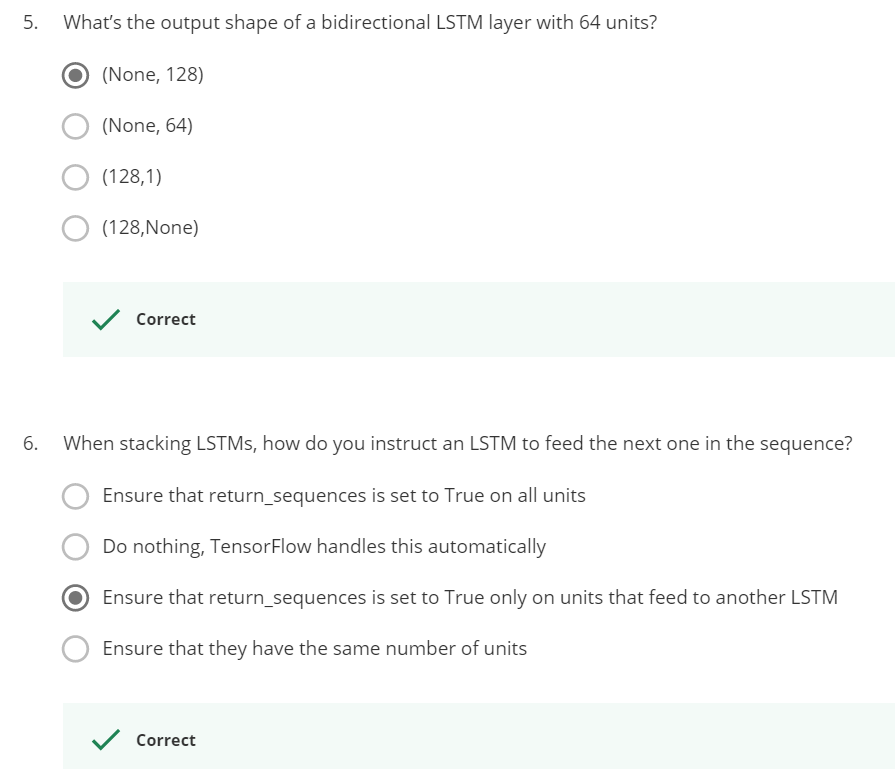


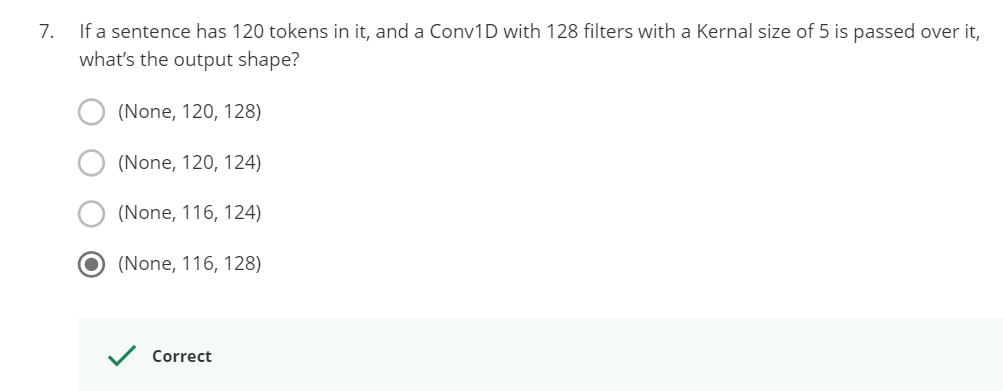
Week 3

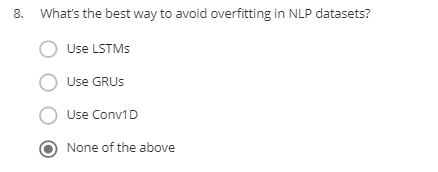












Week 4

