About assignment 4, step 4, please refer to “model_training” folder in the following github repository: https://github.com/lazy-mind/asg4

We made changes in the following files to enable it running local and on sagemaker notebook:

1. "sentiment_dataset.py" (load feature files from s3 directory)
2. "sentiment_model_cnn.py" (load dictionary from s3 directory)
3. "training_config.json" (add parameter for connection)

In “sentiment_dataset.py” file:

(1) Import a new library

```
import os
import json
import math
import tensorflow as tf
import numpy as np
from tensorflow.data import Dataset
import sagemaker
```

(2) read json files, different methods for local loading and cloud loading

```
    if config["cloud"]==0:
        with open(json_path, "r") as file:

            for line in file:

                entry = json.loads(line)

                if len(entry["features"]) != config["padding_size"]:
                    raise ValueError(
                        "The size of the features of the entry with twitterid {} was not expected".format(
                            entry["twitterid"]))

                labels.append(entry["sentiment"] / 4)
                features.append(entry["features"])
    else:
        contents = json_path.split("\n")
        for line in contents:
            entry = json.loads(line)

            if len(entry["features"]) != config["padding_size"]:
                raise ValueError(
                    "The size of the features of the entry with twitterid {} was not expected".format(
                        entry["twitterid"]))

            labels.append(entry["sentiment"] / 4)
            features.append(entry["features"])
```

(3) Load json files, different methods for local loading and cloud loading

```
77      if config["cloud"]==0:
78          all_files = os.listdir(directory)
79          for file in all_files:
80              features, labels = _load_json_file(os.path.join(directory, file), config)
81              all_features += features
82              all_labels += labels
83      else:
84          s = sagemaker.Session()
85          all_files= s.list_s3_files(config["bucket"], directory)
86          for file in all_files[1:]:
87              features, labels = _load_json_file(s.read_s3_file(config["bucket"],file), config)
88              all_features += features
89              all_labels += labels
```

In the "sentiment_model_cnn.py"

(1) Import a new library

```
7   import os
8   import tensorflow as tf
9   import numpy as np
10  import sagemaker
```

(2) Different location to read dictionary, based on running on local or in the cloud

```
25      if config["cloud"]==0:
26          file = open(config["embeddings_path"], "r")
27          for word_vector in file:
28              if(len(word_vector.split())==25):
29                  # print(len(word_vector.split()))
30                  embedding_matrix[index,:] = word_vector.split()[0:]
31              else:
32                  embedding_matrix[index,:] = word_vector.split()[1:]
33              index += 1
34      else:
35          s = sagemaker.Session()
36          all_files= s.list_s3_files(config["bucket"],config["embeddings_path"])
37          result = s.read_s3_file(config["bucket"],all_files[0])
38          file = result.split("\n")[:-1]
39          for word_vector in file:
40              if(len(word_vector.split())==25):
41                  # print(len(word_vector.split()))
42                  embedding_matrix[index,:] = word_vector.split()[0:]
43              else:
44                  embedding_matrix[index,:] = word_vector.split()[1:]
45              index += 1
```

In "training_config.json"

(1) Add 2 parameters:

“bucket”: the s3 bucket that will be connected to
“cloud”: if set to 0, means running on local machine. If set to 1, means running on aws sagemake notebook

```
{
    "embeddings_dictionary_size": 1193516,
    "embeddings_vector_size": 25,
    "padding_size": 140,
    "batch_size": 100,
    "embeddings_path": "dict/glove_vector.txt",
    "input_tensor_name": "embedding_input",
    "embedding_layer_name": "embedding",
    "bucket": "asg4",
    "cloud": 0
}
```

Command for running the code in sagemaker notebook terminal (change green string to the cloud folder name where json features are stored):
python -W ignore sentiment_training.py --train train_features --validation dev_features --eval eval_features --model_output_dir model/ --model_dir model/ --num_epoch 10