TV Script Generation

In this project, you'll generate your own Simpsons TV scripts using RNNs. You'll be using part of the Simpsons dataset of scripts from 27 seasons. The Neural Network you'll build will generate a new TV script for a scene at Moe's Tavern.

Get the Data

The data is already provided for you. You'll be using a subset of the original dataset. It consists of only the scenes in Moe's Tavern. This doesn't include other versions of the tavern, like "Moe's Cavern", "Flaming Moe's", "Uncle Moe's Family Feed-Bag", etc..

In [3]:

```
"""
DON'T MODIFY ANYTHING IN THIS CELL
"""
import helper

data_dir = './data/simpsons/moes_tavern_lines.txt'
text = helper.load_data(data_dir)
# Ignore notice, since we don't use it for analysing the data
text = text[81:]
```

Explore the Data

Play around with `view_sentence_range` to view different parts of the data.

In [4]:

```
view_sentence_range = (0, 10)

"""
DON'T MODIFY ANYTHING IN THIS CELL
"""
import numpy as np

print('Dataset Stats')
print('Roughly the number of unique words: {}'.format(len({word: None for word
in text.split()})))
scenes = text.split('\n\n')
print('Number of scenes: {}'.format(len(scenes)))
sentence_count_scene = [scene.count('\n') for scene in scenes]
print('Average number of sentences in each scene:
{}'.format(np.average(sentence_count_scene)))

sentences = [sentence for scene in scenes for sentence in scene.split('\n')]
print('Number of lines: {}'.format(len(sentences)))
word_count_sentence = [len(sentence.split()) for sentence in sentences]
print('Average number of words in each line:
{}'.format(np.average(word_count_sentence)))

print()
print('The sentences {} to {}:'.format(*view_sentence_range))
print('\n'.join(text.split('\n')[view_sentence_range[0]:view_sentence_range[1]]
))
```

```
Dataset Stats
Roughly the number of unique words: 11492
Number of scenes: 262
Average number of sentences in each scene: 15.251908396946565
Number of lines: 4258
Average number of words in each line: 11.50164396430249

The sentences 0 to 10:

Moe_Szyslak: (INTO PHONE) Moe's Tavern. Where the elite meet to drink.
Bart_Simpson: Eh, yeah, hello, is Mike there? Last name, Rotch.
Moe_Szyslak: (INTO PHONE) Hold on, I'll check. (TO BARFLIES) Mike Rotch. Mike
Rotch. Hey, has anybody seen Mike Rotch, lately?
Moe_Szyslak: (INTO PHONE) Listen you little puke. One of these days I'm gonna
catch you, and I'm gonna carve my name on your back with an ice pick.
Moe_Szyslak: What's the matter Homer? You're not your normal effervescent self.
Homer_Simpson: I got my problems, Moe. Give me another one.
Moe_Szyslak: Homer, hey, you should not drink to forget your problems.
Barney_Gumble: Yeah, you should only drink to enhance your social skills.
```

In [5]:

```
view_sentence_range = (20, 25)

"""
DON'T MODIFY ANYTHING IN THIS CELL
"""
import numpy as np

print('Dataset Stats')
print('Roughly the number of unique words: {}'.format(len({word: None for word
in text.split()})))
scenes = text.split('\n\n')
print('Number of scenes: {}'.format(len(scenes)))
sentence_count_scene = [scene.count('\n') for scene in scenes]
print('Average number of sentences in each scene:
{}'.format(np.average(sentence_count_scene)))

sentences = [sentence for scene in scenes for sentence in scene.split('\n')]
print('Number of lines: {}'.format(len(sentences)))
word_count_sentence = [len(sentence.split()) for sentence in sentences]
print('Average number of words in each line:
{}'.format(np.average(word_count_sentence)))

print()
print('The sentences {} to {}:'.format(*view_sentence_range))
print('\n'.join(text.split('\n')[view_sentence_range[0]:view_sentence_range[1]]
))
```

```
Dataset Stats
Roughly the number of unique words: 11492
Number of scenes: 262
Average number of sentences in each scene: 15.251908396946565
```

```
Number of lines: 4258
Average number of words in each line: 11.50164396430249

The sentences 20 to 25:

Moe_Szyslak: Looks like this is the end.
Barney_Gumble: That's all right. I couldn't have led a richer life.
Barney_Gumble: So the next time somebody tells you county folk are good, honest
people, you can spit in their faces for me!
```

Implement Preprocessing Functions

The first thing to do to any dataset is preprocessing. Implement the following preprocessing functions below:

- Lookup Table
- Tokenize Punctuation

Lookup Table

To create a word embedding, you first need to transform the words to ids. In this function, create two dictionaries:

- Dictionary to go from the words to an id, we'll call `vocab_to_int`
- Dictionary to go from the id to word, we'll call `int_to_vocab`

Return these dictionaries in the following tuple `(vocab_to_int, int_to_vocab)`

In [7]:

```
import numpy as np
import problem_unittests as tests

def create_lookup_tables(text):
    """
    Create lookup tables for vocabulary
    :param text: The text of tv scripts split into words
    :return: A tuple of dicts (vocab_to_int, int_to_vocab)
    """
    # TODO: Implement Function
    vocab_to_int = {word: index for index,word in enumerate(set(text))}
    int_to_vocab = {index: word for index,word in enumerate(set(text))}
    return (vocab_to_int, int_to_vocab)

"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
tests.test_create_lookup_tables(create_lookup_tables)
```

```
Tests Passed
```

Tokenize Punctuation

We'll be splitting the script into a word array using spaces as delimiters. However, punctuations like periods and exclamation marks make it hard for the neural network to distinguish between the word "bye" and "bye!".

Implement the function `token_lookup` to return a dict that will be used to tokenize symbols like "!" into "||Exclamation_Mark||". Create a dictionary for the following symbols where the symbol is the key and value is the token:

- Period (.)
- Comma (,)
- Quotation Mark (")
- Semicolon (;)
- Exclamation mark (!)
- Question mark (?)
- Left Parentheses (( )
- Right Parentheses ( ))
- Dash (--)
- Return (\n)

This dictionary will be used to token the symbols and add the delimiter (space) around it. This separates the symbols as it's own word, making it easier for the neural network to predict on the next word. Make sure you don't use a token that could be confused as a word. Instead of using the token "dash", try using something like "||dash||".

In [8]:

```
def token_lookup():
    """
    Generate a dict to turn punctuation into a token.
    :return: Tokenize dictionary where the key is the punctuation and the value
is the token
    """
    # TODO: Implement Function
    dict_token = {'.' : '||Period||',
                  ',' : '||Comma||',
                  '"' : '||QuotationMark||',
                  ';': '||Semicolon||',
                  '!'  : '||ExclamationMark||',
                  '?' : '||QuestionMark||',
                  '(' : '||LeftParentheses||',
                  ')' : '||RightParentheses||',
                  '--' : '||Dash||',
                  '\n' : '||Return||'}
    return dict_token

"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
tests.test_tokenize(token_lookup)
```

```
Tests Passed
```

Preprocess all the data and save it

Running the code cell below will preprocess all the data and save it to file.

In [9]:

```
"""
DON'T MODIFY ANYTHING IN THIS CELL
"""
# Preprocess Training, Validation, and Testing Data
helper.preprocess_and_save_data(data_dir, token_lookup, create_lookup_tables)
```

Check Point

This is your first checkpoint. If you ever decide to come back to this notebook or have to restart the notebook, you can start from here. The preprocessed data has been saved to disk.

In [10]:

```
"""
DON'T MODIFY ANYTHING IN THIS CELL
"""
import helper
import numpy as np
import problem_unittests as tests

int_text, vocab_to_int, int_to_vocab, token_dict = helper.load_preprocess()
```

Build the Neural Network

You'll build the components necessary to build a RNN by implementing the following functions below:

- get_inputs
- get_init_cell
- get_embed
- build_rnn
- build_nn
- get_batches

Check the Version of TensorFlow and Access to GPU

In [11]:

```
"""
DON'T MODIFY ANYTHING IN THIS CELL
"""
from distutils.version import LooseVersion
import warnings
import tensorflow as tf

# Check TensorFlow Version
assert LooseVersion(tf.__version__) >= LooseVersion('1.0'), 'Please use
TensorFlow version 1.0 or newer'
print('TensorFlow Version: {}'.format(tf.__version__))

# Check for a GPU
if not tf.test.gpu_device_name():
    warnings.warn('No GPU found. Please use a GPU to train your neural
network.')
else:
    print('Default GPU Device: {}'.format(tf.test.gpu_device_name()))
```

```
TensorFlow Version: 1.0.0
C:\Anaconda3\lib\site-packages\ipykernel\__main__.py:14: UserWarning: No GPU
found. Please use a GPU to train your neural network.
```

Input

Implement the `get_inputs()` function to create TF Placeholders for the Neural Network. It should create the following placeholders:

- Input text placeholder named "input" using the TF Placeholder `name` parameter.
- Targets placeholder
- Learning Rate placeholder

Return the placeholders in the following the tuple `(Input, Targets, LearingRate)`

In [12]:

```
def get_inputs():
    """
    Create TF Placeholders for input, targets, and learning rate.
    :return: Tuple (input, targets, learning rate)
    """
    # TODO: Implement Function
    inputs = tf.placeholder(dtype=tf.int32, shape=[None, None], name='input')
    targets = tf.placeholder(dtype=tf.int32, shape=[None, None],
name='targets')
    learning_rate = tf.placeholder(dtype=tf.float32, name='learning_rate')
    return inputs, targets, learning_rate
"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
tests.test_get_inputs(get_inputs)
```

```
Tests Passed
```

Build RNN Cell and Initialize

Stack one or more `BasicLSTMCells` in a `MultiRNNCell`.

- The Rnn size should be set using `rnn_size`
- Initalize Cell State using the MultiRNNCell's `zero_state()` function
- Apply the name "initial_state" to the initial state using `tf.identity()`

Return the cell and initial state in the following tuple `(Cell, InitialState)`

In [19]:

```
def get_init_cell(batch_size, rnn_size, keep_prob=0.7, layers=3):
    """
    Create an RNN Cell and initialize it.
    :param batch_size: Size of batches
    :param rnn_size: Size of RNNs
    :return: Tuple (cell, initialize state)
    """
    # TODO: Implement Function
    lstm = tf.contrib.rnn.BasicLSTMCell(rnn_size)
    drop = tf.contrib.rnn.DropoutWrapper(lstm, output_keep_prob=keep_prob)
    multi = tf.contrib.rnn.MultiRNNCell([drop] * layers)
    init_state = multi.zero_state(batch_size, tf.float32)
    init_state = tf.identity(init_state, 'initial_state')
    return multi, init_state

"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
tests.test_get_init_cell(get_init_cell)
```

```
Tests Passed
```

Word Embedding

Apply embedding to `input_data` using TensorFlow. Return the embedded sequence.

In [20]:

```
def get_embed(input_data, vocab_size, embed_dim):
    """
    Create embedding for <input_data>.
    :param input_data: TF placeholder for text input.
    :param vocab_size: Number of words in vocabulary.
    :param embed_dim: Number of embedding dimensions
    :return: Embedded input.
    """
    # TODO: Implement Function
    embeddings = tf.Variable(tf.random_uniform((vocab_size, embed_dim), -1, 1))
    embed = tf.nn.embedding_lookup(embeddings, input_data)
    return embed

"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
tests.test_get_embed(get_embed)
```

```
Tests Passed
```

Build RNN

You created a RNN Cell in the `get_init_cell()` function. Time to use the cell to create a RNN.

- Build the RNN using the `tf.nn.dynamic_rnn()`
- Apply the name "final_state" to the final state using `tf.identity()`

Return the outputs and final_state state in the following tuple `(Outputs, FinalState)`

In [21]:

```
def build_rnn(cell, inputs):
    """
    Create a RNN using a RNN Cell
    :param cell: RNN Cell
    :param inputs: Input text data
    :return: Tuple (Outputs, Final State)
    """
    # TODO: Implement Function
    Outputs, final_state = tf.nn.dynamic_rnn(cell,inputs,None,dtype=tf.float32)
    FinalState = tf.identity(final_state,name="final_state")
    return (Outputs,FinalState)

"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
tests.test_build_rnn(build_rnn)
```

```
Tests Passed
```

Build the Neural Network

Apply the functions you implemented above to:

- Apply embedding to `input_data` using your `get_embed(input_data, vocab_size, embed_dim)` function.
- Build RNN using `cell` and your `build_rnn(cell, inputs)` function.
- Apply a fully connected layer with a linear activation and `vocab_size` as the number of outputs.

Return the logits and final state in the following tuple (Logits, FinalState)

In [22]:

```
def build_nn(cell, rnn_size, input_data, vocab_size):
    """
    Build part of the neural network
    :param cell: RNN cell
    :param rnn_size: Size of rnns
    :param input_data: Input data
    :param vocab_size: Vocabulary size
    :return: Tuple (Logits, FinalState)
    """
    # TODO: Implement Function
    embed = get_embed(input_data, vocab_size, rnn_size)
    outputs, final_state = build_rnn(cell, embed)
    logits = tf.contrib.layers.fully_connected(outputs, vocab_size,
activation_fn=None)
    return logits, final_state

"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
tests.test_build_nn(build_nn)
```

```
Tests Passed
```

Batches

Implement `get_batches` to create batches of input and targets using `int_text`. The batches should be a Numpy array with the shape `(number of batches, 2, batch size, sequence length)`. Each batch contains two elements:

- The first element is a single batch of **input** with the shape `[batch size, sequence length]`
- The second element is a single batch of **targets** with the shape `[batch size, sequence length]`

If you can't fill the last batch with enough data, drop the last batch.

For exmple, `get_batches([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15], 2, 3)` would return a Numpy array of the following:

```
[
  # First Batch
  [
    # Batch of Input
    [[ 1  2  3], [ 7  8  9]],
    # Batch of targets
    [[ 2  3  4], [ 8  9 10]]
  ],
```

```
      # Second Batch
      [
        # Batch of Input
        [[ 4  5  6], [10 11 12]],
        # Batch of targets
        [[ 5  6  7], [11 12 13]]
      ]
    ]
```

In [24]:

```
def get_batches(int_text, batch_size, seq_length):
    """
    Return batches of input and target
    :param int_text: Text with the words replaced by their ids
    :param batch_size: The size of batch
    :param seq_length: The length of sequence
    :return: Batches as a Numpy array
    """
    # TODO: Implement Function
    n_batches = len(int_text) // (batch_size * seq_length)
    output = []
    for i in range(n_batches):
        inputs = []
        targets = []
        for j in range(batch_size):
            ndx = i * seq_length + j * seq_length
            inputs.append(int_text[ndx:ndx + seq_length])
            targets.append(int_text[ndx + 1:ndx + seq_length + 1])
        output.append([inputs, targets])
    return np.array(output)


"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
tests.test_get_batches(get_batches)
```

```
Tests Passed
```

Neural Network Training

Hyperparameters

Tune the following parameters:

- Set `num_epochs` to the number of epochs.
- Set `batch_size` to the batch size.
- Set `rnn_size` to the size of the RNNs.
- Set `seq_length` to the length of sequence.
- Set `learning_rate` to the learning rate.
- Set `show_every_n_batches` to the number of batches the neural network should print progress.

In [40]:

```
# Number of Epochs
num_epochs = 60
# Batch Size
batch_size = 86
# RNN Size
rnn_size = 256
# Sequence Length
seq_length = 25
# Learning Rate
learning_rate = 0.01
# Show stats for every n number of batches
show_every_n_batches = get_batches(int_text, batch_size, seq_length).shape[0]
"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
save_dir = './save'
```

Build the Graph

Build the graph using the neural network you implemented.

In [41]:

```
"""
DON'T MODIFY ANYTHING IN THIS CELL
"""
from tensorflow.contrib import seq2seq

train_graph = tf.Graph()
with train_graph.as_default():
    vocab_size = len(int_to_vocab)
    input_text, targets, lr = get_inputs()
    input_data_shape = tf.shape(input_text)
    cell, initial_state = get_init_cell(input_data_shape[0], rnn_size)
    logits, final_state = build_nn(cell, rnn_size, input_text, vocab_size)

    # Probabilities for generating words
    probs = tf.nn.softmax(logits, name='probs')

    # Loss function
    cost = seq2seq.sequence_loss(
        logits,
        targets,
        tf.ones([input_data_shape[0], input_data_shape[1]]))

    # Optimizer
    optimizer = tf.train.AdamOptimizer(lr)

    # Gradient Clipping
    gradients = optimizer.compute_gradients(cost)
    capped_gradients = [(tf.clip_by_value(grad, -1., 1.), var) for grad, var in
gradients]
    train_op = optimizer.apply_gradients(capped_gradients)
```

Train

Train the neural network on the preprocessed data. If you have a hard time getting a good loss, check the forms to see if anyone is having the same problem.

In [42]:

```
"""
DON'T MODIFY ANYTHING IN THIS CELL
"""
batches = get_batches(int_text, batch_size, seq_length)

with tf.Session(graph=train_graph) as sess:
    sess.run(tf.global_variables_initializer())

    for epoch_i in range(num_epochs):
        state = sess.run(initial_state, {input_text: batches[0][0]})

        for batch_i, (x, y) in enumerate(batches):
            feed = {
                input_text: x,
                targets: y,
                initial_state: state,
                lr: learning_rate}
            train_loss, state, _ = sess.run([cost, final_state, train_op],
feed)
            # Show every <show_every_n_batches> batches
            if (epoch_i * len(batches) + batch_i) % show_every_n_batches == 0:
                print('Epoch {:>3} Batch {:>4}/{}   train_loss =
{:.3f}'.format(
                    epoch_i,
                    batch_i,
                    len(batches),
                    train_loss))

    # Save Model
    saver = tf.train.Saver()
    saver.save(sess, save_dir)
    print('Model Trained and Saved')
```

```
Epoch   0 Batch    0/32   train_loss = 8.822
Epoch   1 Batch    0/32   train_loss = 5.607
……
Epoch  59 Batch    0/32   train_loss = 0.060
Model Trained and Saved
```

Save Parameters

Save `seq_length` and `save_dir` for generating a new TV script.

In [43]:

```
"""
DON'T MODIFY ANYTHING IN THIS CELL
"""
# Save parameters for checkpoint
helper.save_params((seq_length, save_dir))
```

Checkpoint

In [44]:

```
"""
DON'T MODIFY ANYTHING IN THIS CELL
"""
import tensorflow as tf
import numpy as np
import helper
import problem_unittests as tests

_, vocab_to_int, int_to_vocab, token_dict = helper.load_preprocess()
seq_length, load_dir = helper.load_params()
```

Implement Generate Functions

Get Tensors

Get tensors from `loaded_graph` using the function `get_tensor_by_name()`. Get the tensors using the following names:

- "input:0"
- "initial_state:0"
- "final_state:0"
- "probs:0"

Return the tensors in the following tuple `(InputTensor, InitialStateTensor, FinalStateTensor, ProbsTensor)`

In [45]:

```
def get_tensors(loaded_graph):
    """
    Get input, initial state, final state, and probabilities tensor from <loaded_graph>
    :param loaded_graph: TensorFlow graph loaded from file
    :return: Tuple (InputTensor, InitialStateTensor, FinalStateTensor, ProbsTensor)
    """
    # TODO: Implement Function
    inputs = loaded_graph.get_tensor_by_name('input:0')
    init_state = loaded_graph.get_tensor_by_name('initial_state:0')
    final_state = loaded_graph.get_tensor_by_name('final_state:0')
    probs = loaded_graph.get_tensor_by_name('probs:0')
    return inputs, init_state, final_state, probs


"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
tests.test_get_tensors(get_tensors)
```

```
Tests Passed
```

Choose Word

Implement the `pick_word()` function to select the next word using `probabilities`.

In [46]:

```
def pick_word(probabilities, int_to_vocab):
    """
    Pick the next word in the generated text
    :param probabilities: Probabilites of the next word
    :param int_to_vocab: Dictionary of word ids as the keys and words as the
values
    :return: String of the predicted word
    """
    # TODO: Implement Function
    return int_to_vocab[np.argmax(probabilities)]


"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
tests.test_pick_word(pick_word)
```

```
Tests Passed
```

Generate TV Script

This will generate the TV script for you. Set `gen_length` to the length of TV script you want to generate.

In [47]:

```
gen_length = 200
# homer_simpson, moe_szyslak, or Barney_Gumble
prime_word = 'moe_szyslak'

"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
loaded_graph = tf.Graph()
with tf.Session(graph=loaded_graph) as sess:
    # Load saved model
    loader = tf.train.import_meta_graph(load_dir + '.meta')
    loader.restore(sess, load_dir)

    # Get Tensors from loaded model
    input_text, initial_state, final_state, probs = get_tensors(loaded_graph)

    # Sentences generation setup
    gen_sentences = [prime_word + ':']
    prev_state = sess.run(initial_state, {input_text: np.array([[1]])})

    # Generate sentences
    for n in range(gen_length):
        # Dynamic Input
        dyn_input = [[vocab_to_int[word] for word in gen_sentences[-
seq_length:]]]
        dyn_seq_length = len(dyn_input[0])
```

```
        # Get Prediction
        probabilities, prev_state = sess.run(
            [probs, final_state],
            {input_text: dyn_input, initial_state: prev_state})

        pred_word = pick_word(probabilities[dyn_seq_length-1], int_to_vocab)

        gen_sentences.append(pred_word)

    # Remove tokens
    tv_script = ' '.join(gen_sentences)
    for key, token in token_dict.items():
        ending = ' ' if key in ['\n', '(', '"'] else ''
        tv_script = tv_script.replace(' ' + token.lower(), key)
    tv_script = tv_script.replace('\n ', '\n')
    tv_script = tv_script.replace('( ', '(')

    print(tv_script)
```

```
moe_szyslak: what's the matter homer? you're not your normal effervescent self.
homer_simpson: you're smithers, bad.
moe_szyslak: uh, let me happy power out of things.
little_man:(a us. do you easy,.
moe_szyslak: hey, what's the beer.
moe_szyslak: hey, but my tester's beer.
barney_gumble: i'll check.
.(conspiratorial) let moe_szyslak: take as your me for the wife lot of bars
once. i think aerosmith monkeyshines.
moe_szyslak:(conspiratorial) treat. people, i wrote it to be beer... on? one
instrument once i call it on my much a the wow, what's the you gin is" flaming
moe" is sorry, yeah. do, don't flaming homer".
moe_szyslak: sorry, not is from big excuse me, i couldn't tester's lot your guys
of sure he on the it" flaming moe"?
homer_simpson: hey, bet in a treat.". do know". moe_szyslak:(sobs) i on that
that.
homer_simpson:(for) barkeep
```

The TV Script is Nonsensical

It's ok if the TV script doesn't make any sense. We trained on less than a megabyte of text. In order to get good results, you'll have to use a smaller vocabulary or get more data. Luckly there's more data! As we mentioned in the begging of this project, this is a subset of another dataset. We didn't have you train on all the data, because that would take too long. However, you are free to train your neural network on all the data. After you complete the project, of course.

Helper.py

```
import os
import pickle


def load_data(path):
    """
    Load Dataset from File
    """
    input_file = os.path.join(path)
    with open(input_file, "r") as f:
        data = f.read()

    return data


def preprocess_and_save_data(dataset_path, token_lookup, create_lookup_tables):
    """
    Preprocess Text Data
    """
    text = load_data(dataset_path)

    # Ignore notice, since we don't use it for analysing the data
    text = text[81:]

    token_dict = token_lookup()
    for key, token in token_dict.items():
        text = text.replace(key, ' {} '.format(token))

    text = text.lower()
    text = text.split()

    vocab_to_int, int_to_vocab = create_lookup_tables(text)
    int_text = [vocab_to_int[word] for word in text]
    pickle.dump((int_text, vocab_to_int, int_to_vocab, token_dict),
open('preprocess.p', 'wb'))


def load_preprocess():
    """
    Load the Preprocessed Training data and return them in batches of
<batch_size> or less
    """
    return pickle.load(open('preprocess.p', mode='rb'))


def save_params(params):
    """
    Save parameters to file
    """
    pickle.dump(params, open('params.p', 'wb'))


def load_params():
    """
    Load parameters from file
    """
    return pickle.load(open('params.p', mode='rb'))
```