data pre-process:

1. text ← original text file
2. char2idx ← {char: index} dictionary
3. text_as_int ← [18 47 56 57 ...] text file represented by integer.
4. char_dataset = tf.data.Dataset.from_tensor_slices(text_as_int)

char_dataset ↓ tensor 格式的 text_as_int.

from_tensor_slices. example:

```
dataset = tf.data.Dataset.from_tensor_slices([1, 2, 3])
for element in dataset:
    print(element)
tf.Tensor(1, shape=(), dtype=int32)
:        2        ...
:        3        ...
```

5. sequences = char_dataset.batch(seq_length+1, drop_remainder=True)

sequences. 相当于把 [c1, c2, c3 ... cn]

↓

[[c1, c2, ..., cm] m = batch
  [cm+1, cm+2, ... c2m] here m = seq_length+1
  ⋮
  cn]] m = 100

6. sequences 's duplicated, processed and splited to 2 parts, [Input data, Target_data]

↓

dataset.

7. dataset : [Input_data, Target_data],

where

Input_data = sequences,

Target_data = sequences removed first char

Then:

8. dataset = dataset.shuffle(10000).batch(64)

shape(dataset): (((64, 100), (64, 100)), ...)

Then dataset 's used in training

model.fit(dataset, epochs, callbacks)

dataset 分为 training 和 validation 两部分