In [ ]:

```
#@title Licensed under the Apache License, Version 2.0 (the "Licens
# you may not use this file except in compliance with the License.
# You may obtain a copy of the License at
#
# https://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, softwa
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or i
# See the License for the specific language governing permissions a
# limitations under the License.
```

In [ ]:

```
!wget --no-check-certificate \
    https://storage.googleapis.com/laurencemoroney-blog.appspot.com
    -O /tmp/bbc-text.csv


import csv
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences


# Stopwords list from https://github.com/Yoast/YoastSEO.js/blob/dev
# Convert it to a Python list and paste it here
#we're gonna filter out stop words, since they occur in abundance a
#meaning - as they provide little to no information
stopwords = [ "a", "about", "above", "after", "again", "against", "
```

```
--2020-09-18 05:48:20--  https://storage.googleapis.com/laurencemor
pspot.com/bbc-text.csv (https://storage.googleapis.com/laurencemoro
spot.com/bbc-text.csv)
Resolving storage.googleapis.com (storage.googleapis.com)... 64.233
4.233.183.128, 172.217.212.128, ...
Connecting to storage.googleapis.com (storage.googleapis.com)|64.23
43... connected.
HTTP request sent, awaiting response... 200 OK
Length: 5057493 (4.8M) [application/octet-stream]
Saving to: ‘/tmp/bbc-text.csv’

/tmp/bbc-text.csv   100%[===================>]   4.82M  --.-KB/s

2020-09-18 05:48:21 (107 MB/s) - ‘/tmp/bbc-text.csv’ saved [5057493
```

In [ ]:

```
sentences = []
labels = []
with open("/tmp/bbc-text.csv", 'r') as csvfile:
    # Your Code here
    reader = csv.reader(csvfile, delimiter=',')
    next(reader) #ignore header
    for data_row in reader:
      #labels, sentences
      sentence_to_remove_stop_words = data_row[1]
      labels.append(data_row[0])
      for word in stopwords:
        sentence_to_remove_stop_words = sentence_to_remove_stop_wor
        #sentences have 2 blanks at times
        sentence_to_remove_stop_words = sentence_to_remove_stop_wor
      sentences.append(sentence_to_remove_stop_words)


#successfully separated the labels and the sentences
print(len(sentences))
print(sentences[0])

#Expected output
# 2225
# tv future hands viewers home theatre systems plasma high-definiti
```

```
2225
tv future hands viewers home theatre systems plasma high-definitior
video recorders moving living room way people watch tv will radical
five years time. according expert panel gathered annual consumer el
ow las vegas discuss new technologies will impact one favourite pas
ading trend programmes content will delivered viewers via home netw
atellite telecoms companies broadband service providers front rooms
vices. one talked-about technologies ces digital personal video rec
pvr). set-top boxes like us s tivo uk s sky+ system allow people re
lay pause forward wind tv programmes want. essentially technology a
ersonalised tv. also built-in high-definition tv sets big business
wer take off europe lack high-definition programming. not can peopl
nd adverts can also forget abiding network channel schedules puttir
-la-carte entertainment. us networks cable satellite companies worr
rms advertising revenues well brand identity viewer loyalty channel
us leads technology moment also concern raised europe particularly
ke services like sky+. happens today will see nine months years tim
me bbc broadcast s futurologist told bbc news website. likes bbc nc
advertising revenue yet. pressing issue moment commercial uk broadc
loyalty important everyone. will talking content brands rather netw
aid tim hanlon brand communications firm starcom mediavest. reality
onnections anybody can producer content. added: challenge now hard
ramme much choice. means said stacey jolna senior vice president tv
oup way people find content want watch simplified tv viewers. means
terms channels take leaf google s book search engine future insteac
elp people find want watch. kind channel model might work younger i
on used taking control gadgets play them. might not suit everyone p
sed. older generations comfortable familiar schedules channel branc
ng. perhaps not want much choice put hands mr hanlon suggested. enc
iapers pushing buttons already - everything possible available saic
```

```
ultimately consumer will tell market want. 50 000 new gadgets techr
cased ces many enhancing tv-watching experience. high-definition tv
here many new models lcd (liquid crystal display) tvs launched dvr
uilt instead external boxes. one example launched show humax s 26-i
0-hour tivo dvr dvd recorder. one us s biggest satellite tv compani
even launched branded dvr show 100-hours recording capability insta
arch function. set can pause rewind tv 90 hours. microsoft chief bi
ounced pre-show keynote speech partnership tivo called tivotogo mea
n play recorded programmes windows pcs mobile devices. reflect incr
freeing multimedia people can watch want want.
```

In [ ]:

```python
#instantiate tokenizer with an out-of-vocabulary token
tokenizer = Tokenizer(oov_token="<OOV>")
#tokenize the words in our sentences input
tokenizer.fit_on_texts(sentences)
#create the dictionary of words: encoded value
word_index = tokenizer.word_index
print(len(word_index))
# Expected output
# 29714
```

```
29714
```

In [ ]:

```python
#add all the tokenized sentences into an array
sequences = tokenizer.texts_to_sequences(sentences)
#zero-pad the sentences to create uniformity of length (based on ma
#in our list of sequences)
padded = pad_sequences(sequences, padding="post")
print(padded[0])
print(padded.shape)

# Expected output
# [  96  176 1158 ...    0    0    0]
# (2225, 2442)
```

```
[  96  176 1158 ...    0    0    0]
(2225, 2442)
```

In [ ]:

```
# Your Code Here
#we tokenized the sentences, now we tokenize the labels associating
label_tokenizer = Tokenizer()
label_tokenizer.fit_on_texts(labels)
label_word_index = label_tokenizer.word_index
label_seq = label_tokenizer.texts_to_sequences(labels)
print(label_seq)
print(label_word_index)

# Expected Output
# [[4], [2], [1], [1], [5], [3], [3], [1], [1], [5], [5], [2], [2],
# {'sport': 1, 'business': 2, 'politics': 3, 'tech': 4, 'entertainm
```

```
[[4], [2], [1], [1], [5], [3], [3], [1], [1], [5], [5], [2], [2], [
[2], [3], [1], [2], [4], [4], [4], [1], [1], [4], [1], [5], [4], [3
[4], [5], [5], [2], [3], [4], [5], [3], [2], [3], [1], [2], [1], [4
[3], [3], [2], [1], [3], [2], [2], [1], [3], [2], [1], [1], [2], [2
[1], [2], [4], [2], [5], [4], [2], [3], [2], [3], [1], [2], [4], [2
[2], [2], [1], [3], [2], [5], [3], [3], [2], [5], [2], [1], [1], [3
[1], [2], [1], [2], [5], [5], [1], [2], [3], [3], [4], [1], [5], [1
[5], [1], [5], [1], [5], [5], [3], [1], [1], [5], [3], [2], [4], [2
[1], [3], [1], [4], [5], [1], [2], [2], [4], [5], [4], [1], [2], [2
[1], [4], [2], [1], [5], [1], [4], [1], [4], [3], [2], [4], [5], [1
[2], [5], [3], [3], [5], [3], [2], [5], [3], [3], [5], [3], [1], [2
[2], [5], [1], [2], [2], [1], [4], [1], [4], [4], [1], [2], [1], [3
[2], [3], [2], [4], [3], [5], [3], [4], [2], [1], [2], [1], [4], [5
[3], [5], [1], [5], [3], [1], [5], [1], [1], [5], [1], [3], [3], [5
[3], [2], [5], [4], [1], [4], [1], [5], [3], [1], [5], [4], [2], [4
[4], [2], [1], [2], [1], [2], [1], [5], [2], [2], [5], [1], [1], [3
[3], [3], [4], [1], [4], [3], [2], [4], [5], [4], [1], [1], [2], [2
[4], [1], [5], [1], [3], [4], [5], [2], [1], [5], [1], [4], [3], [4
[3], [3], [1], [2], [4], [5], [3], [4], [2], [5], [1], [5], [1], [5
[1], [2], [1], [1], [5], [1], [3], [3], [2], [5], [4], [2], [1], [2
[2], [2], [3], [2], [3], [5], [5], [2], [1], [2], [3], [2], [4], [5
[1], [5], [2], [2], [3], [4], [5], [4], [3], [2], [1], [3], [2], [5
[4], [3], [1], [5], [2], [3], [2], [2], [3], [1], [4], [2], [2], [5
[1], [2], [5], [4], [4], [5], [5], [5], [3], [1], [3], [4], [2], [5
[5], [3], [3], [1], [1], [2], [3], [5], [2], [1], [2], [2], [1], [2
[3], [1], [4], [4], [2], [4], [1], [5], [2], [3], [2], [5], [2], [3
[2], [4], [2], [1], [1], [2], [1], [1], [5], [1], [1], [1], [4], [2
[3], [1], [1], [2], [4], [2], [3], [1], [3], [4], [2], [1], [5], [2
[2], [1], [2], [3], [2], [2], [1], [5], [4], [3], [4], [2], [1], [2
[4], [2], [1], [1], [5], [3], [3], [3], [1], [3], [4], [4], [5], [3
[2], [1], [1], [4], [2], [1], [1], [3], [1], [1], [2], [1], [5], [4
[3], [4], [2], [2], [2], [4], [2], [2], [1], [1], [1], [1], [2], [4
[1], [4], [2], [4], [5], [3], [1], [2], [3], [2], [4], [4], [3], [4
[2], [5], [1], [3], [5], [1], [1], [3], [4], [5], [4], [1], [3], [2
[2], [5], [1], [1], [4], [3], [5], [3], [5], [3], [4], [3], [5], [1
[5], [1], [5], [4], [2], [1], [3], [5], [3], [5], [5], [5], [3], [5
[4], [4], [1], [1], [4], [4], [1], [5], [5], [1], [4], [5], [1], [1
[3], [4], [2], [1], [5], [1], [5], [3], [4], [5], [5], [2], [5], [5
[4], [3], [1], [4], [1], [3], [3], [5], [4], [2], [4], [4], [4], [2
[1], [4], [2], [2], [5], [5], [1], [4], [2], [4], [5], [1], [4], [3
[2], [3], [3], [2], [1], [4], [1], [4], [3], [5], [4], [1], [5], [4
[5], [1], [4], [1], [1], [3], [5], [2], [3], [5], [2], [2], [4], [2
[1], [4], [3], [4], [3], [2], [3], [5], [1], [2], [2], [2], [5], [1
```

```
[5], [1], [5], [3], [3], [3], [1], [1], [1], [4], [3], [1], [3], [3
[1], [2], [5], [1], [2], [2], [4], [2], [5], [5], [5], [2], [5], [5
[2], [1], [4], [1], [1], [3], [2], [1], [4], [2], [1], [4], [1], [1
[2], [1], [2], [4], [3], [4], [2], [1], [1], [2], [2], [2], [2], [3
[4], [2], [1], [3], [2], [4], [2], [1], [2], [3], [5], [1], [2], [3
[2], [2], [2], [1], [3], [5], [1], [3], [1], [3], [3], [2], [2], [1
[1], [5], [2], [2], [2], [4], [1], [4], [3], [4], [4], [4], [1], [4
[5], [4], [1], [5], [4], [1], [1], [2], [5], [4], [2], [1], [2], [3
[4], [2], [3], [2], [4], [1], [2], [5], [2], [3], [1], [5], [3], [1
[3], [3], [1], [5], [5], [2], [2], [1], [4], [4], [1], [5], [4], [4
[5], [4], [1], [1], [2], [5], [2], [2], [2], [5], [1], [5], [4], [4
[4], [4], [5], [5], [1], [1], [3], [2], [5], [1], [3], [5], [4], [3
[2], [5], [3], [4], [3], [3], [1], [3], [3], [5], [4], [1], [3], [1
[2], [2], [3], [1], [1], [1], [5], [4], [4], [2], [5], [1], [3], [4
[4], [4], [2], [2], [1], [2], [2], [4], [3], [5], [2], [2], [2], [2
[1], [3], [4], [4], [2], [2], [5], [3], [5], [1], [4], [1], [5], [1
[2], [1], [3], [3], [5], [2], [1], [3], [3], [1], [5], [3], [2], [4
[2], [2], [5], [5], [4], [4], [2], [2], [5], [1], [2], [5], [4], [4
[1], [1], [1], [3], [3], [1], [3], [1], [2], [5], [1], [4], [5], [1
[2], [4], [4], [1], [5], [1], [5], [1], [5], [3], [5], [5], [4], [5
[3], [1], [3], [4], [2], [3], [1], [3], [1], [5], [1], [3], [1], [1
[1], [3], [1], [1], [2], [4], [5], [3], [4], [5], [3], [5], [3], [5
[5], [3], [5], [5], [4], [4], [1], [1], [5], [5], [4], [5], [3], [4
[4], [1], [2], [5], [5], [4], [5], [4], [2], [5], [1], [5], [2], [1
[3], [4], [5], [3], [2], [5], [5], [3], [2], [5], [1], [3], [1], [2
[2], [2], [5], [4], [1], [5], [5], [2], [1], [4], [4], [5], [1], [2
[3], [2], [2], [5], [3], [2], [2], [4], [3], [1], [4], [5], [3], [2
[5], [3], [4], [2], [2], [3], [2], [1], [5], [1], [5], [4], [3], [2
[2], [2], [1], [2], [4], [5], [3], [2], [3], [2], [1], [4], [2], [3
[2], [5], [1], [3], [3], [1], [3], [2], [4], [5], [1], [1], [4], [2
[4], [1], [3], [1], [2], [2], [2], [3], [5], [1], [3], [4], [2], [2
[5], [4], [4], [1], [1], [5], [4], [5], [1], [3], [4], [2], [1], [5
[5], [1], [2], [1], [4], [3], [3], [4], [5], [3], [5], [2], [2], [3
[1], [1], [1], [3], [2], [1], [2], [4], [1], [2], [2], [1], [3], [4
[4], [1], [1], [2], [2], [2], [2], [3], [5], [4], [2], [2], [1], [2
[5], [1], [3], [2], [2], [4], [5], [2], [2], [2], [3], [2], [3], [4
[5], [1], [4], [3], [2], [4], [1], [2], [2], [5], [4], [2], [2], [1
[1], [3], [1], [2], [1], [2], [3], [3], [2], [3], [4], [5], [1], [2
[3], [3], [4], [5], [2], [3], [3], [1], [4], [2], [1], [5], [1], [5
[1], [3], [5], [4], [2], [1], [3], [4], [1], [5], [2], [1], [5], [1
[4], [3], [1], [2], [5], [4], [4], [3], [4], [5], [4], [1], [2], [4
[1], [4], [3], [3], [3], [3], [5], [5], [5], [2], [3], [3], [1], [1
[3], [2], [2], [4], [1], [4], [2], [4], [3], [3], [1], [2], [3], [1
[2], [2], [5], [5], [1], [2], [4], [4], [3], [2], [3], [1], [5], [5
[2], [2], [4], [4], [1], [1], [3], [4], [1], [4], [2], [1], [2], [3
[2], [4], [3], [5], [4], [2], [1], [5], [4], [4], [5], [3], [4], [5
[1], [1], [1], [3], [4], [1], [2], [1], [1], [2], [4], [1], [2], [5
[1], [3], [4], [5], [3], [1], [3], [4], [2], [5], [1], [3], [2], [4
[3], [2], [1], [3], [5], [4], [5], [1], [4], [2], [3], [5], [4], [3
[2], [5], [2], [2], [3], [2], [2], [3], [4], [5], [3], [5], [5], [2
[3], [5], [1], [5], [3], [5], [5], [5], [2], [1], [3], [1], [5], [4
[3], [5], [2], [1], [2], [3], [3], [2], [1], [4], [4], [4], [2], [3
[1], [1], [5], [2], [1], [1], [3], [3], [3], [5], [3], [2], [4], [2
[5], [2], [1], [3], [5], [1], [5], [3], [3], [2], [3], [1], [5], [5
[4], [4], [3], [4], [2], [4], [1], [1], [5], [2], [4], [5], [2], [4
[5], [5], [3], [3], [1], [2], [2], [4], [5], [1], [3], [2], [4], [5
[5], [3], [3], [4], [1], [3], [2], [3], [5], [4], [1], [3], [5], [5
```

```
[4], [4], [1], [5], [4], [3], [4], [1], [3], [3], [1], [5], [1], [3
[5], [1], [5], [2], [2], [5], [5], [5], [4], [1], [2], [2], [3], [3
[5], [1], [1], [4], [3], [1], [2], [1], [2], [4], [1], [1], [2], [5
[4], [1], [2], [3], [2], [5], [4], [5], [3], [2], [5], [3], [5], [3
[1], [1], [1], [4], [4], [1], [3], [5], [4], [1], [5], [2], [5], [3
[4], [2], [1], [3], [2], [5], [5], [5], [3], [5], [3], [5], [1], [5
[3], [2], [3], [4], [1], [4], [1], [2], [3], [4], [5], [5], [3], [5
[1], [3], [2], [4], [1], [3], [3], [5], [1], [3], [3], [2], [4], [4
[1], [1], [2], [3], [2], [4], [1], [4], [3], [5], [1], [2], [1], [5
[1], [3], [1], [2], [1], [2], [1], [1], [5], [5], [2], [4], [4], [2
[2], [1], [1], [3], [1], [4], [1], [4], [1], [1], [2], [2], [4], [1
[4], [3], [1], [2], [5], [5], [4], [3], [1], [1], [4], [2], [4], [5
[3], [2], [5], [1], [5], [5], [2], [1], [3], [4], [2], [1], [5], [4
[1], [1], [2], [2], [2], [2], [2], [5], [2], [3], [3], [4], [4], [5
[2], [3], [1], [1], [2], [4], [2], [4], [1], [2], [2], [3], [1], [1
[5], [5], [3], [2], [3], [3], [2], [4], [3], [3], [3], [3], [3], [5
[3], [1], [3], [1], [4], [1], [1], [1], [5], [4], [5], [4], [1], [4
[5], [5], [2], [5], [5], [3], [2], [1], [4], [4], [3], [2], [1], [2
[3], [5], [1], [1], [2], [3], [4], [4], [2], [2], [1], [3], [5], [1
[5], [4], [1], [5], [2], [3], [1], [3], [4], [5], [1], [3], [2], [5
[3], [1], [3], [2], [2], [3], [2], [4], [1], [2], [5], [2], [1], [1
[3], [4], [3], [3], [1], [1], [1], [2], [4], [5], [2], [1], [2], [1
[2], [2], [2], [2], [1], [1], [1], [2], [2], [5], [2], [2], [2], [1
[4], [2], [1], [1], [1], [2], [5], [4], [4], [4], [3], [2], [2], [4
[1], [1], [3], [3], [3], [1], [1], [3], [3], [4], [2], [1], [1], [1
[1], [2], [2], [2], [2], [1], [3], [1], [4], [4], [1], [4], [2], [5
[2], [4], [4], [3], [5], [2], [5], [2], [4], [3], [5], [3], [5], [5
[4], [4], [2], [3], [1], [5], [2], [3], [5], [2], [4], [1], [4], [3
[2], [3], [3], [2], [2], [2], [4], [3], [2], [3], [2], [5], [3], [1
[1], [5], [4], [4], [2], [4], [1], [2], [2], [3], [1], [4], [4], [4
[1], [3], [2], [3], [3], [5], [4], [2], [4], [1], [5], [5], [1], [2
[4], [1], [5], [2], [3], [3], [3], [4], [4], [2], [3], [2], [3], [3
[4], [2], [4], [5], [4], [4], [1], [3], [1], [1], [3], [5], [5], [2
[1], [2], [2], [4], [2], [4], [4], [1], [2], [3], [1], [2], [2], [1
[4], [5], [1], [1], [5], [2], [4], [1], [1], [3], [4], [2], [3], [1
[5], [4], [4], [4], [2], [1], [5], [5], [4], [2], [3], [4], [1], [1
[3], [2], [1], [5], [5], [1], [5], [4], [4], [2], [2], [2], [1], [1
[2], [4], [2], [2], [1], [2], [3], [2], [2], [4], [2], [4], [3], [4
[4], [5], [1], [3], [5], [2], [4], [2], [4], [5], [4], [1], [2], [2
[3], [1]]
{'sport': 1, 'business': 2, 'politics': 3, 'tech': 4, 'entertainmer
```