Ingest

```
import datetime
from packaging import version
from collections import Counter
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# TensorFlow and tf.keras
import tensorflow as tf
from tensorflow import keras
import tensorflow_datasets as tfds
```

```
%matplotlib inline
np.set_printoptions(precision=3, suppress=True)
```

```
def plot_graphs(history, metric):
  plt.plot(history.history[metric])
  plt.plot(history.history['val_'+metric], '')
  plt.xlabel("Epochs")
  plt.ylabel(metric)
  plt.legend([metric, 'val_'+metric])
```

```
#register  ag_news_subset so that tfds.load doesn't generate a checksum (mismatch)
!python -m tensorflow_datasets.scripts.download_and_prepare --register_checksums --

# https://www.tensorflow.org/datasets/splits
# The full `train` and `test` splits, interleaved together.
ri = tfds.core.ReadInstruction('train') + tfds.core.ReadInstruction('test')
dataset_all, info = tfds.load('ag_news_subset', with_info=True,  split=ri, as_super
```

```
    2021-08-09 03:11:37.088574: I tensorflow/stream_executor/platform/default/dso_
    I0809 03:11:39.267617 139995537213312 download_and_prepare.py:200] Running dov
    ag_news_subset
    2021-08-09 03:11:39.276826: I tensorflow/core/platform/cloud/google_auth_prov
    2021-08-09 03:11:39.495077: I tensorflow/core/platform/cloud/google_auth_prov
    2021-08-09 03:11:39.695174: I tensorflow/core/platform/cloud/google_auth_prov
    I0809 03:11:39.884630 139995537213312 dataset_info.py:434] Load pre-computed I
    2021-08-09 03:11:39.895594: I tensorflow/core/platform/cloud/google_auth_prov
    2021-08-09 03:11:40.448568: I tensorflow/core/platform/cloud/google_auth_prov
    I0809 03:11:40.833542 139995537213312 dataset_info.py:361] Load dataset info 
    I0809 03:11:40.835463 139995537213312 download_and_prepare.py:138] download_a
    I0809 03:11:40.835853 139995537213312 dataset_builder.py:357] Generating data
    Downloading and preparing dataset ag_news_subset/1.0.0 (download: 11.24 MiB, 
    Dl Completed...:   0% 0/1 [00:04<?, ? url/s]
    Dl Size...: 0 MiB [00:04, ? MiB/s]

    Extraction completed...: 0 file [00:04, ? file/s]
    Dl Completed...:   0% 0/1 [00:04<?, ? url/s]
    Dl Size...: 1 MiB [00:04,  4.34s/ MiB]

    Dl Completed...:   0% 0/1 [00:04<?, ? url/s]
    Dl Size...: 2 MiB [00:04,  4.34s/ MiB]

    Dl Completed...:   0% 0/1 [00:04<?, ? url/s]
    Dl Size...: 3 MiB [00:04,  4.34s/ MiB]

    Dl Completed...:   0% 0/1 [00:04<?, ? url/s]
    Dl Size...: 4 MiB [00:04,  4.34s/ MiB]

    Dl Completed...:   0% 0/1 [00:04<?, ? url/s]
    Dl Size...: 5 MiB [00:04,  4.34s/ MiB]

    Dl Completed...:   0% 0/1 [00:04<?, ? url/s]
    Dl Size...: 6 MiB [00:04,  4.34s/ MiB]

    Dl Completed...:   0% 0/1 [00:04<?, ? url/s]
    Dl Size...: 7 MiB [00:04,  4.34s/ MiB]

    Dl Completed...:   0% 0/1 [00:04<?, ? url/s]
    Dl Size...: 8 MiB [00:04,  4.34s/ MiB]

    Dl Completed...:   0% 0/1 [00:04<?, ? url/s]
    Dl Size...: 9 MiB [00:04,  4.34s/ MiB]

    Dl Completed...:   0% 0/1 [00:04<?, ? url/s]
    Dl Size...: 10 MiB [00:04,  4.34s/ MiB]

    Dl Completed...:   0% 0/1 [00:04<?, ? url/s]
    Dl Size...: 11 MiB [00:04,  4.34s/ MiB]
```

```
tfds.as_dataframe(dataset_all.take(10),info)
```

	description	label
0	AMD #39;s new dual-core Opteron chip is designed mainly for corporate computing applications, including databases, Web services, and financial transactions.	3 (Sci/Tech)
1	Reuters - Major League Baseball\Monday announced a decision on the appeal filed by Chicago Cubs\pitcher Kerry Wood regarding a suspension stemming from an\incident earlier this season.	1 (Sports)

EDA

```
# classes dictionary
categories =dict(enumerate(info.features["label"].names))
categories
train_categories = [categories[label] for label in dataset_all.map(lambda text, lab
Counter(train_categories).most_common()
```

```
[('Sci/Tech', 31900), ('Sports', 31900), ('Business', 31900), ('World', 31900
```

```
encoder = tf.keras.layers.experimental.preprocessing.TextVectorization(max_tokens=
encoder.adapt(dataset_all.map(lambda text, label: text))
vocab = np.array(encoder.get_vocabulary())
```

```
print(f"There are {len(vocab)} vocabulary words in the corpus.")
```

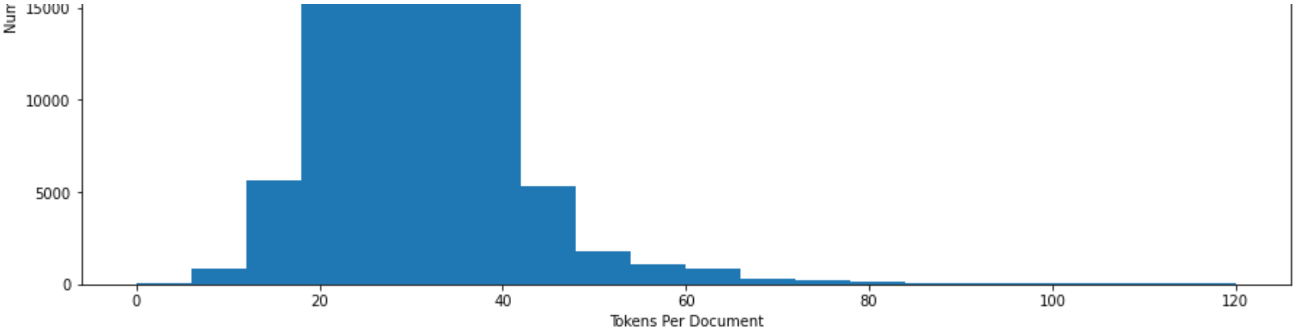
```
There are 95976 vocabulary words in the corpus.
```

```
%%time
doc_sizes = []
corpus = []
for example, _ in dataset_all.as_numpy_iterator():
  enc_example = encoder(example)
  doc_sizes.append(len(enc_example))
  corpus+=list(enc_example.numpy())
```

```
CPU times: user 11min 44s, sys: 1min 3s, total: 12min 47s
Wall time: 10min 21s
```

```
print(f"There are {len(corpus)} words in the corpus of {len(doc_sizes)} news articl
print(f"Each news article has between {min(doc_sizes)} and {max(doc_sizes)} tokens
```

```
There are 3909695 words in the corpus of 127600 news articles.
```



Tokenizing on Base Setup (1000 tokens)

```
%%time
encoder_1000 = tf.keras.layers.experimental.preprocessing.TextVectorization(max_tok
encoder_1000.adapt(dataset_all.map(lambda text, label: text))
vocab_1000 = np.array(encoder_1000.get_vocabulary());
```

```
    CPU times: user 1min 36s, sys: 10.7 s, total: 1min 47s
    Wall time: 1min 22s
```

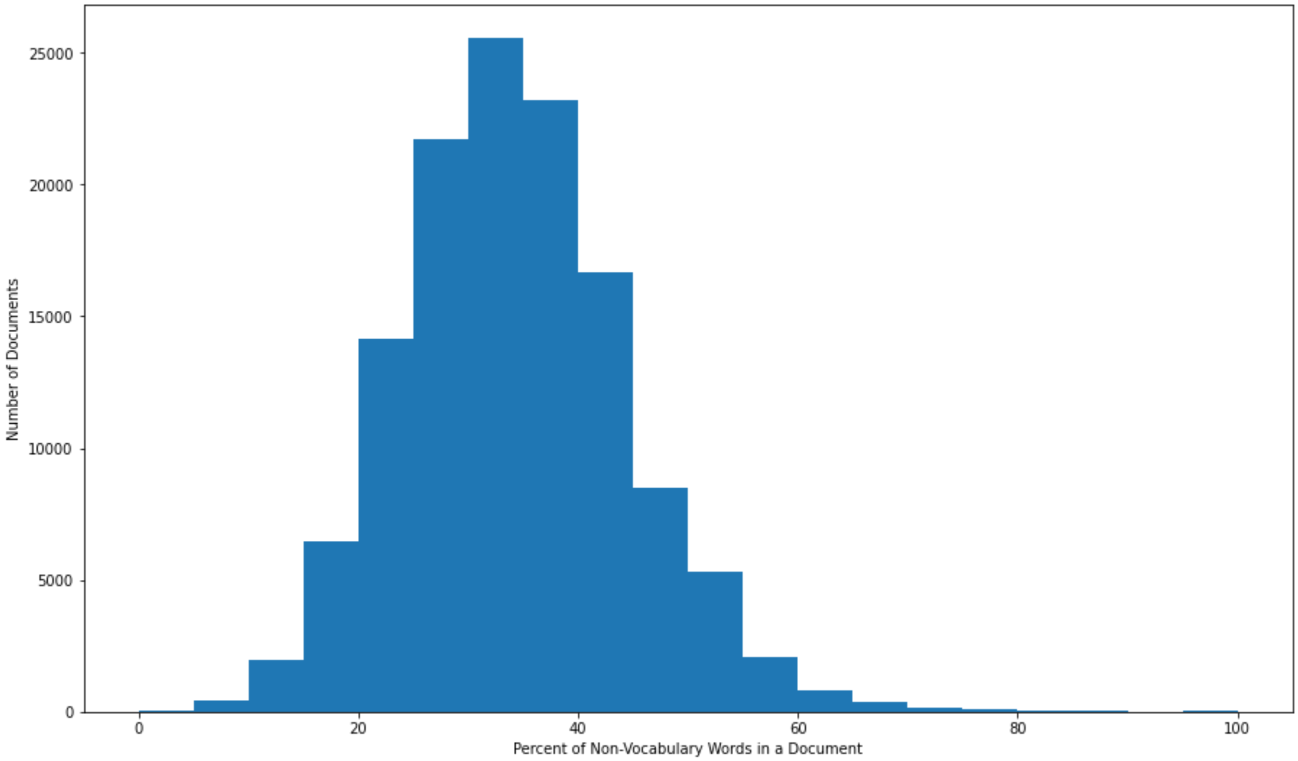
```
vocab_1000[:20]
```

```
    "by" is in the vocabulary.
    "the" is in the vocabulary.
    "hat" is *not* in the vocabulary.
    "on" is in the vocabulary.
    "the" is in the vocabulary.
    "ground." is *not* in the vocabulary.
```

```
%%time
doc1000_sizes = []
corpus1000 = []
count1000=0
useless = 0
# stop = 0
percents = []
for example, _ in dataset_all.as_numpy_iterator():
  # stop+=1
  # if stop > 5: break
  enc_example = encoder_1000(example)
  num_ones = tf.math.count_nonzero(enc_example==1).numpy()
  percent_ones = round(num_ones*100/len(enc_example))
  # print(f"{percent_ones}%")
  percents.append(percent_ones)

  s = set(list(enc_example.numpy()))
  if s == {1}: useless+=1
```

```
plt.ylabel('Number of Documents')
plt.xlabel('Percent of Non-Vocabulary Words in a Document');
```



```
citation: "@misc{zhang2015characterlevel,\n    title={Character-level Convolu
location {
  urls: "https://arxiv.org/abs/1509.01626"
}
splits {
  name: "test"
  shard_lengths: 7600
  num_bytes: 2226751
}
splits {
  name: "train"
  shard_lengths: 120000
  num_bytes: 35301386
}
supervised_keys {
```

```
labels:  [1 1 1]
texts:  [b'The United Nations Security Council is traveling to Nairobi, Kenya
 b' CHICAGO (Reuters) - Wal-Mart Stores Inc. &lt;A HREF="http://www.investor.
 b" BAGHDAD (Reuters) - Twin car bombs detonated outside a  police station nea

labels:  [0 2 0]
```