News-*ish*

A Machine Learning Classification Model

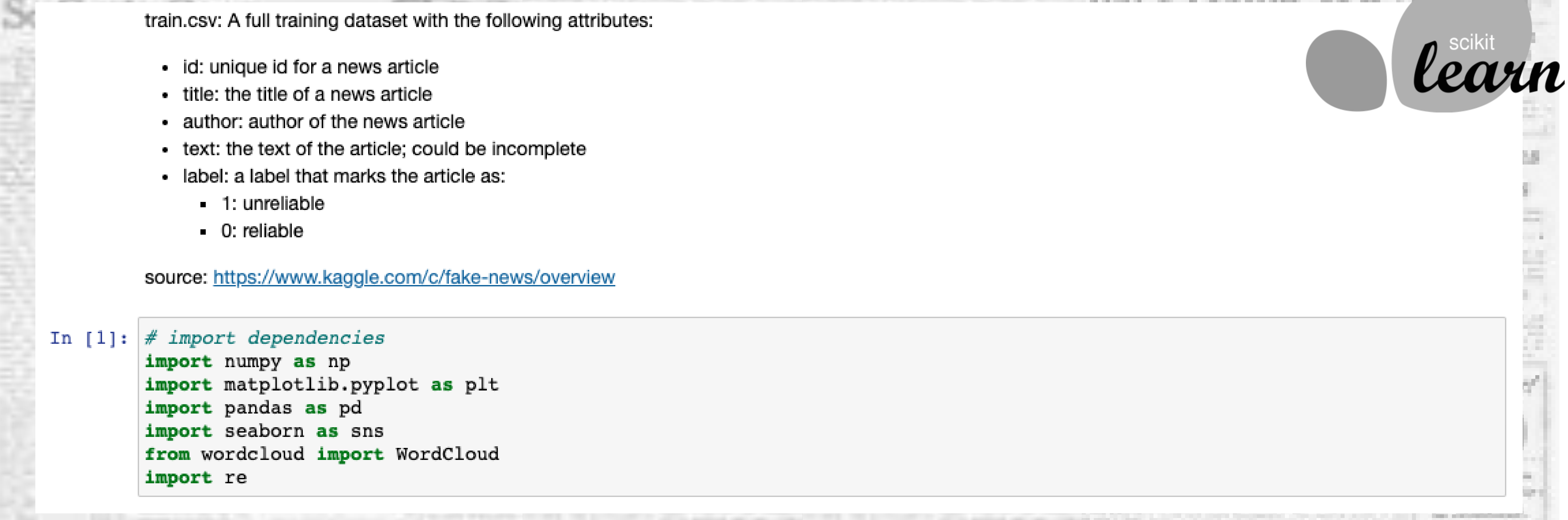
by Matt Mead, Allan Hunt, Alex Mogren, and Emilio Bello

Project Description

This project is a machine learning exercise in which we used an NLP classification model to determine whether a news article is considered a reliable source of information. The model implements a Multinomial Naïve Bayes algorithm to predict these news sources as “reliable” or “unreliable”.

The process started by exploring the data, preprocessing the model input, training the model and testing it. Finally, we developed a website that features a user input form on our home page in which you can paste your own article to determine whether it is a reliable information source.

Data Exploration



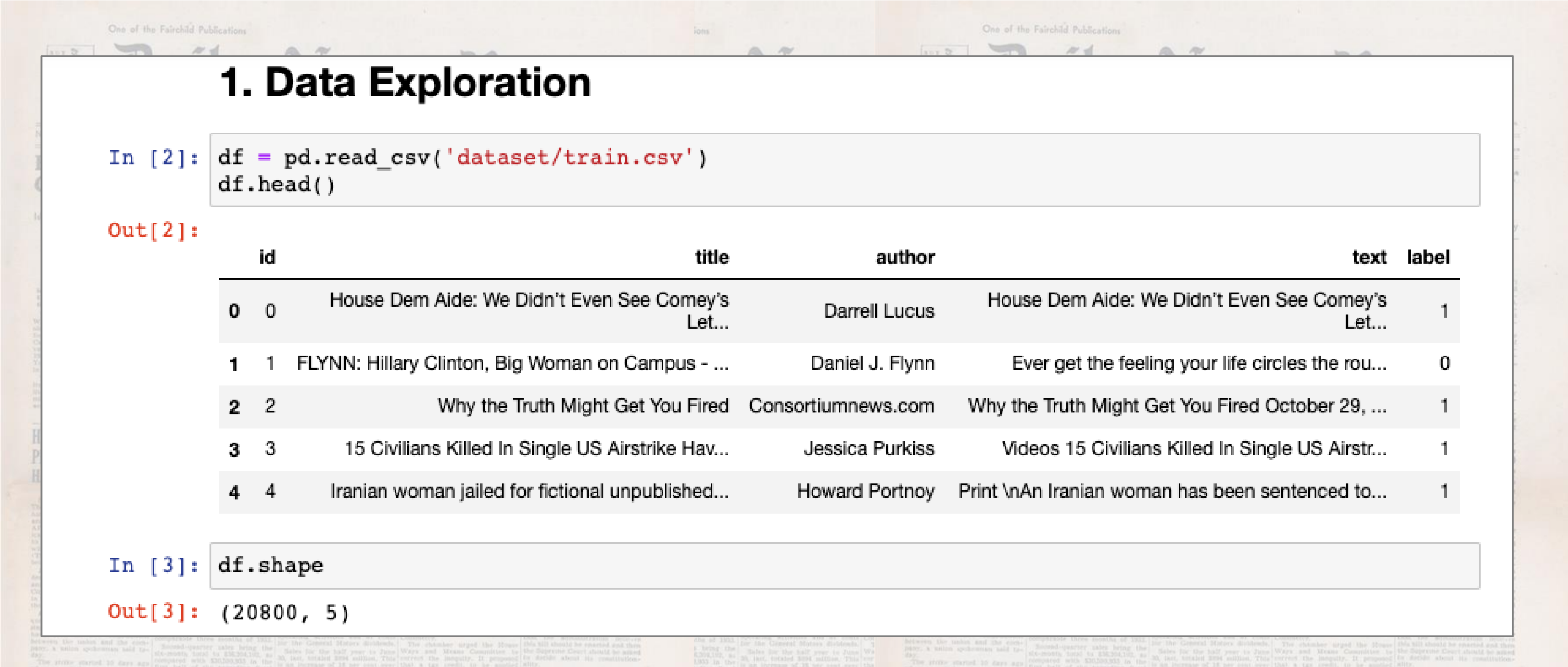
train.csv: A full training dataset with the following attributes:

- id: unique id for a news article
- title: the title of a news article
- author: author of the news article
- text: the text of the article; could be incomplete
- label: a label that marks the article as:
  - 1: unreliable
  - 0: reliable

source: https://www.kaggle.com/c/fake-news/overview

In [1]:
```
# import dependencies
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
from wordcloud import WordCloud
import re
```

Data Exploration



1. Data Exploration

In [2]:
```
df = pd.read_csv('dataset/train.csv')
df.head()
```

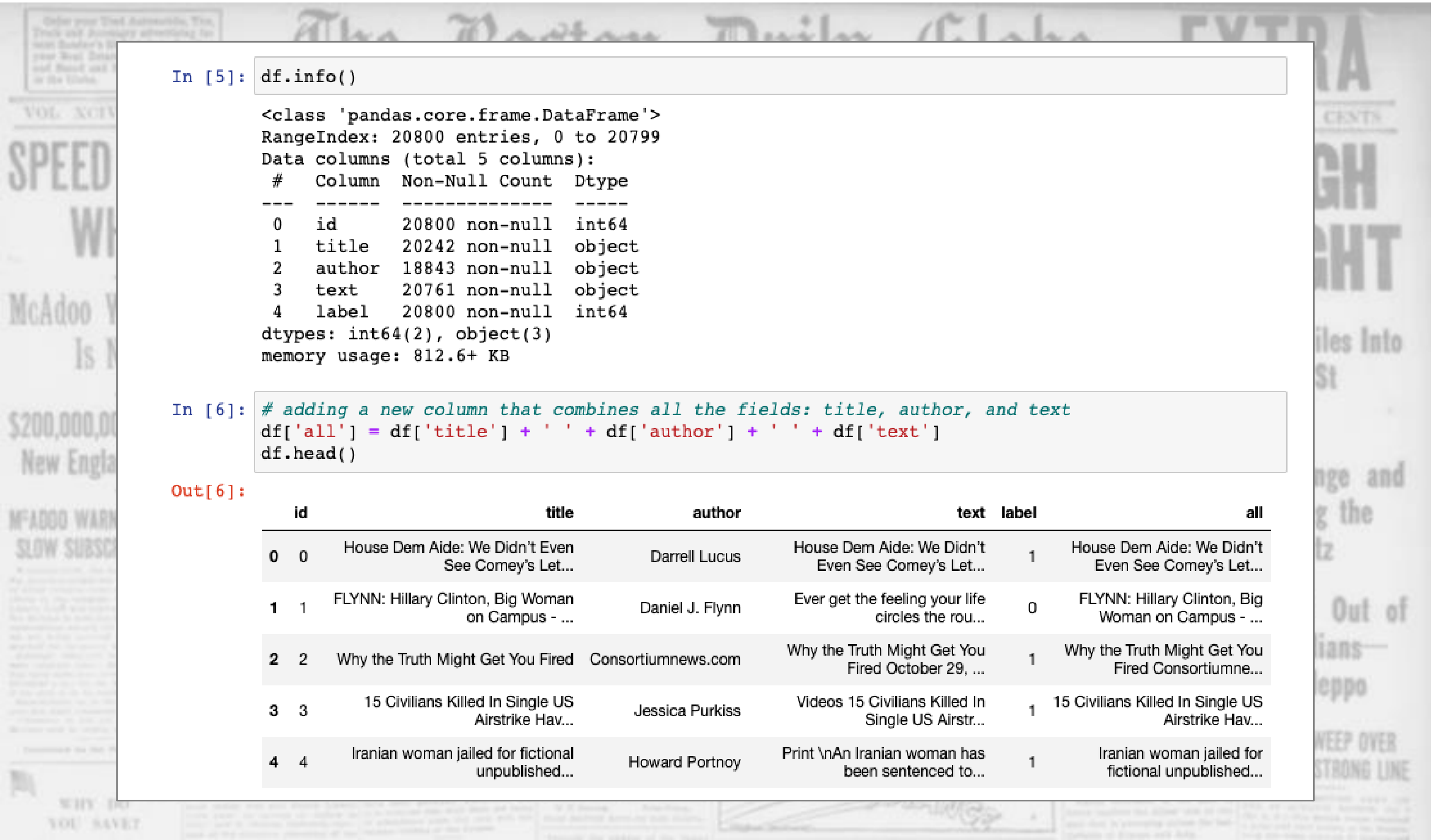
Out[2]:

	id	title	author	text	label
0	0	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

In [3]:
```
df.shape
```

Out[3]:
```
(20800, 5)
```

Data Exploration



```
In [5]: df.info()
```

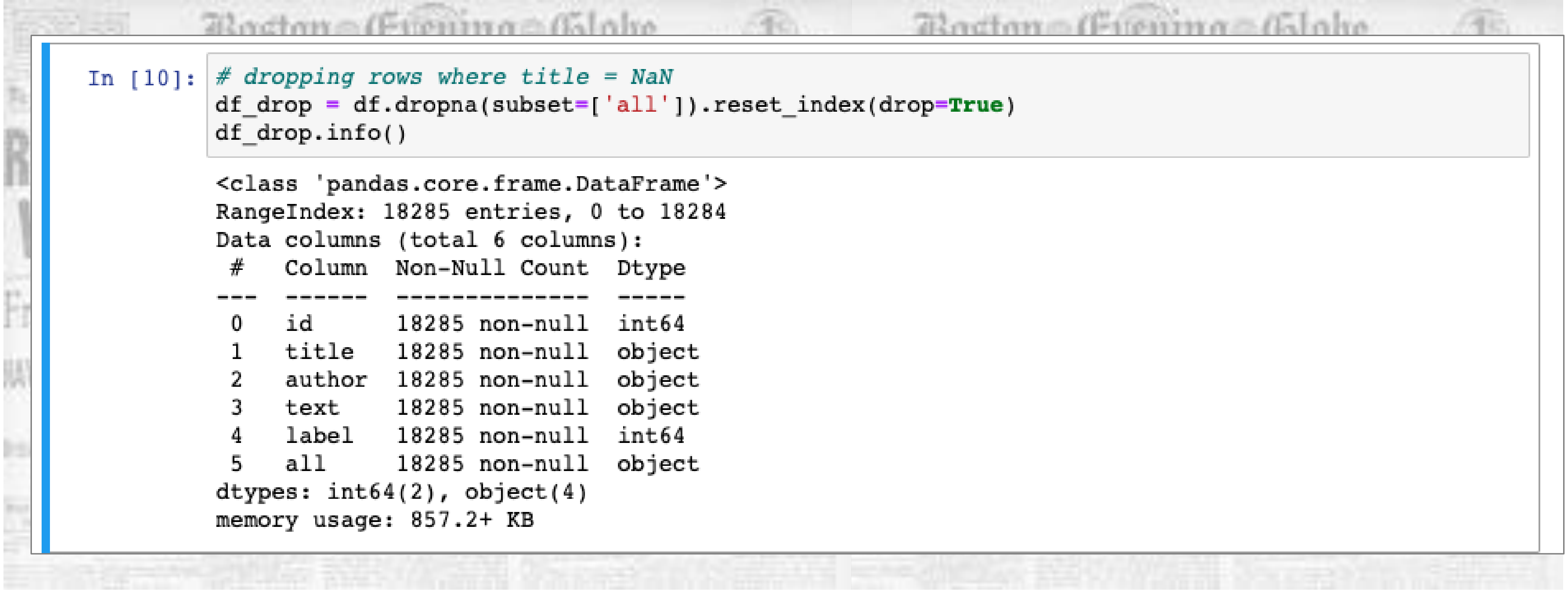
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20800 entries, 0 to 20799
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      20800 non-null  int64
 1   title   20242 non-null  object
 2   author  18843 non-null  object
 3   text    20761 non-null  object
 4   label   20800 non-null  int64
dtypes: int64(2), object(3)
memory usage: 812.6+ KB
```

```
In [6]: # adding a new column that combines all the fields: title, author, and text
        df['all'] = df['title'] + ' ' + df['author'] + ' ' + df['text']
        df.head()
```

Out[6]:

	id	title	author	text	label	all
0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	1	House Dem Aide: We Didn’t Even See Comey’s Let...
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0	FLYNN: Hillary Clinton, Big Woman on Campus - ...
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1	Why the Truth Might Get You Fired Consortiumne...
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1	15 Civilians Killed In Single US Airstrike Hav...
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1	Iranian woman jailed for fictional unpublished...

Data Exploration

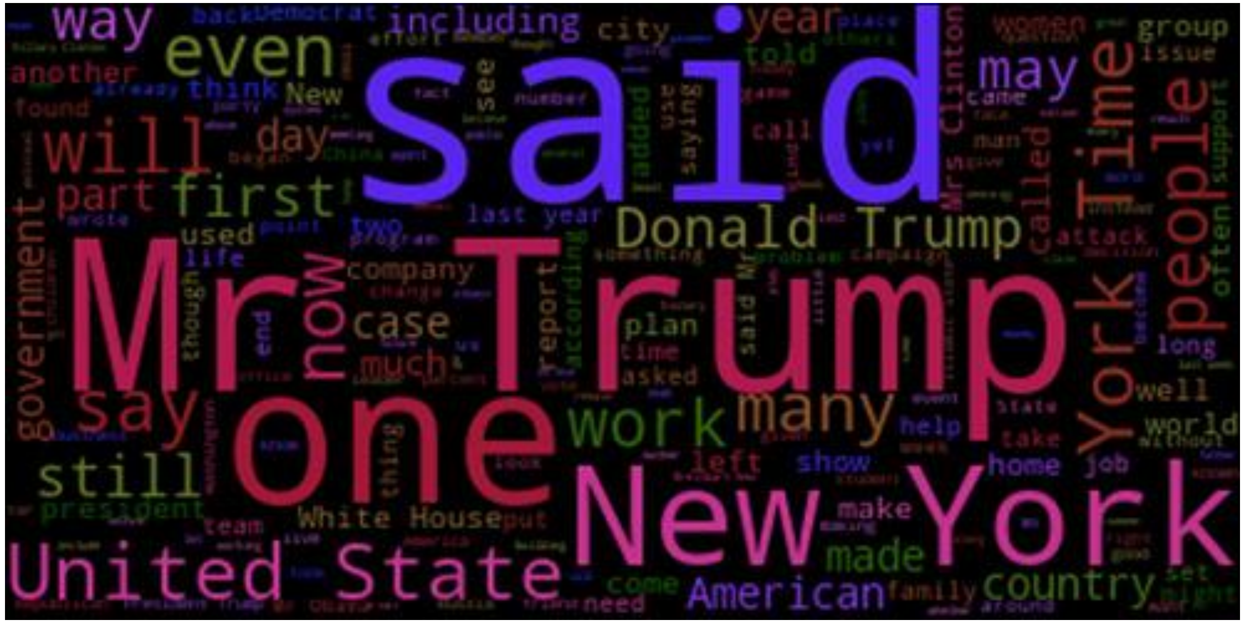


```
In [10]: # dropping rows where title = NaN
         df_drop = df.dropna(subset=['all']).reset_index(drop=True)
         df_drop.info()
```

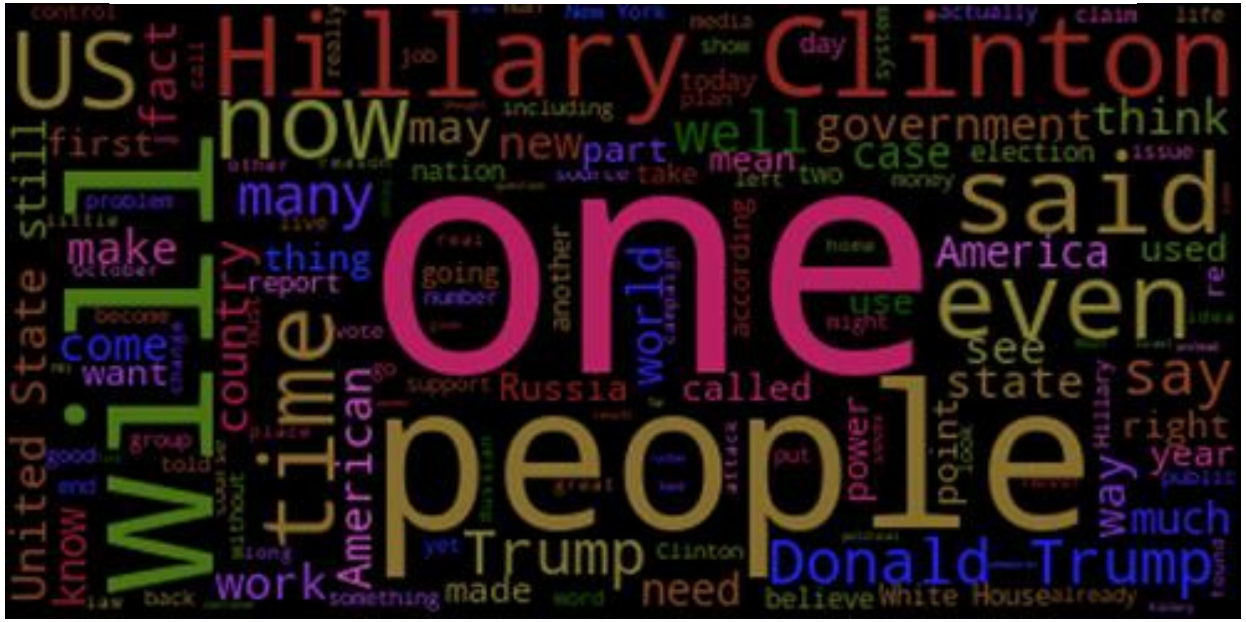
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18285 entries, 0 to 18284
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   id      18285 non-null  int64 
 1   title   18285 non-null  object
 2   author  18285 non-null  object
 3   text    18285 non-null  object
 4   label   18285 non-null  int64 
 5   all     18285 non-null  object
dtypes: int64(2), object(4)
memory usage: 857.2+ KB
```

Data Exploration

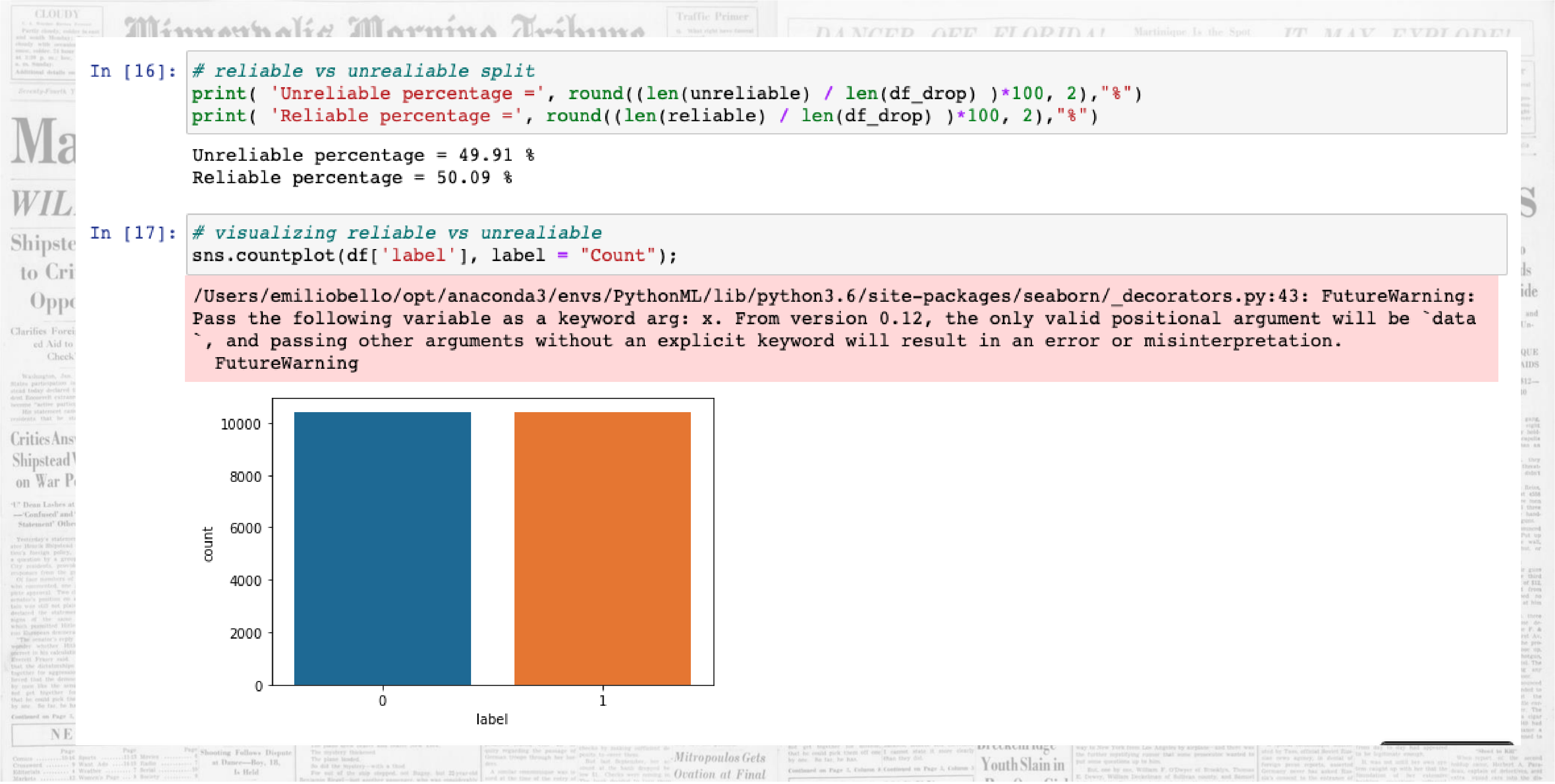
Reliable



UnReliable



Data Exploration



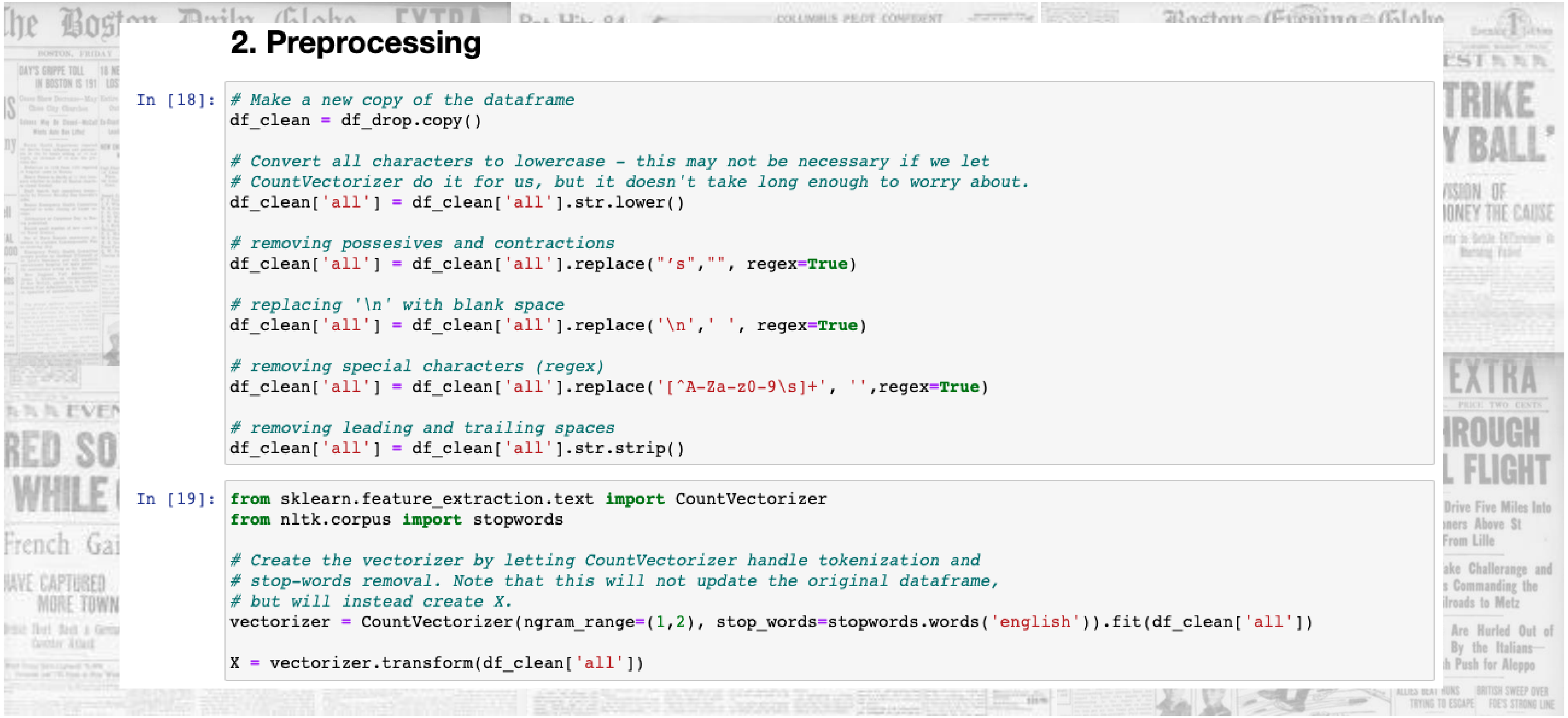
In [16]:
```
# reliable vs unrealiable split
print( 'Unreliable percentage =', round((len(unreliable) / len(df_drop) )*100, 2),"%")
print( 'Reliable percentage =', round((len(reliable) / len(df_drop) )*100, 2),"%")
```

```
Unreliable percentage = 49.91 %
Reliable percentage = 50.09 %
```

In [17]:
```
# visualizing reliable vs unrealiable
sns.countplot(df['label'], label = "Count");
```

```
/Users/emiliobello/opt/anaconda3/envs/PythonML/lib/python3.6/site-packages/seaborn/_decorators.py:43: FutureWarning:
Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data
`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
```

Preprocessing

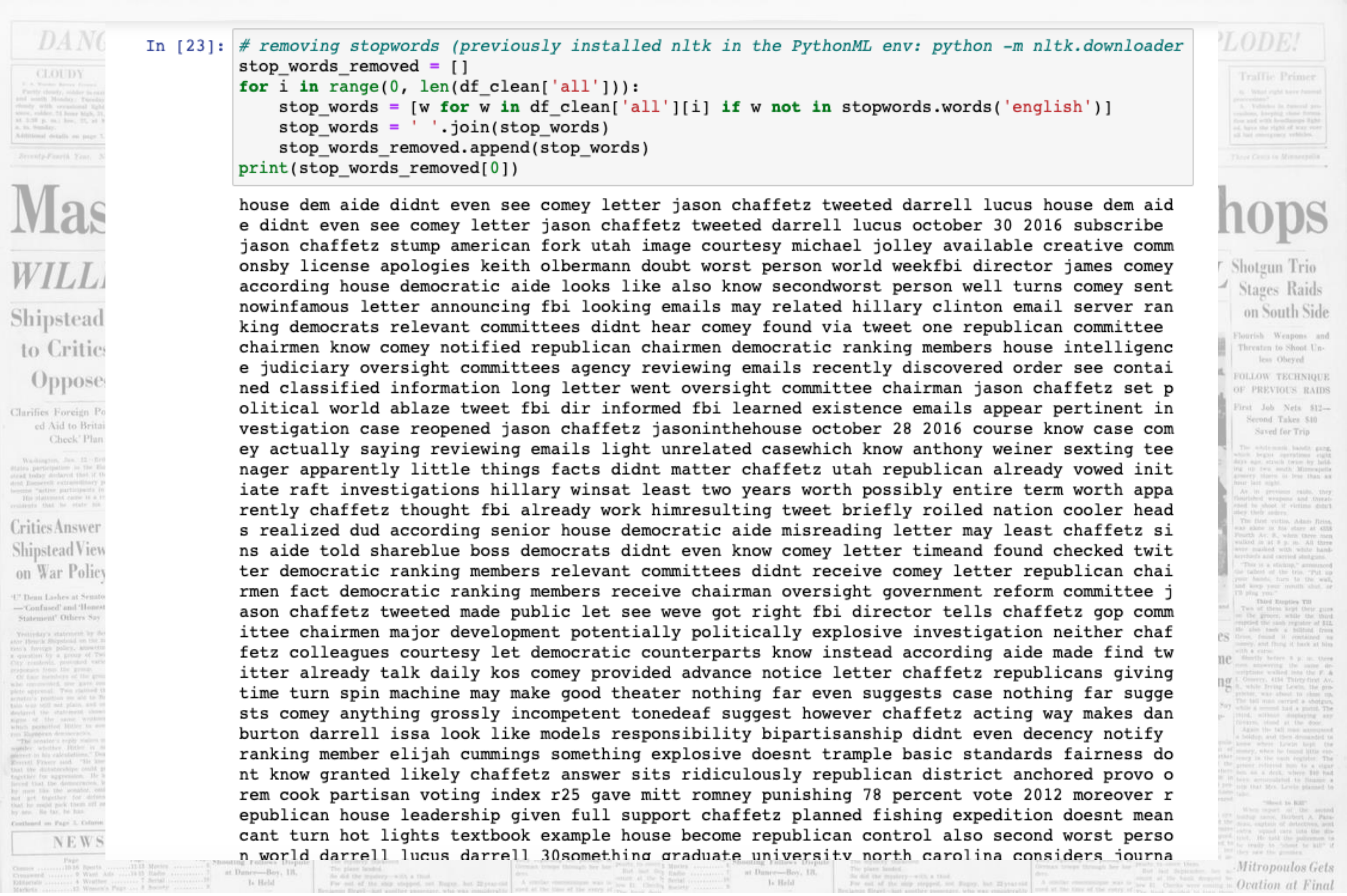


2. Preprocessing

In [18]:
```
# Make a new copy of the dataframe
df_clean = df_drop.copy()

# Convert all characters to lowercase - this may not be necessary if we let
# CountVectorizer do it for us, but it doesn't take long enough to worry about.
df_clean['all'] = df_clean['all'].str.lower()

# removing possesives and contractions
df_clean['all'] = df_clean['all'].replace("’s","", regex=True)

# replacing '\n' with blank space
df_clean['all'] = df_clean['all'].replace('\n',' ', regex=True)

# removing special characters (regex)
df_clean['all'] = df_clean['all'].replace('[^A-Za-z0-9\s]+', '',regex=True)

# removing leading and trailing spaces
df_clean['all'] = df_clean['all'].str.strip()
```

In [19]:
```
from sklearn.feature_extraction.text import CountVectorizer
from nltk.corpus import stopwords

# Create the vectorizer by letting CountVectorizer handle tokenization and
# stop-words removal. Note that this will not update the original dataframe,
# but will instead create X.
vectorizer = CountVectorizer(ngram_range=(1,2), stop_words=stopwords.words('english')).fit(df_clean['all'])

X = vectorizer.transform(df_clean['all'])
```

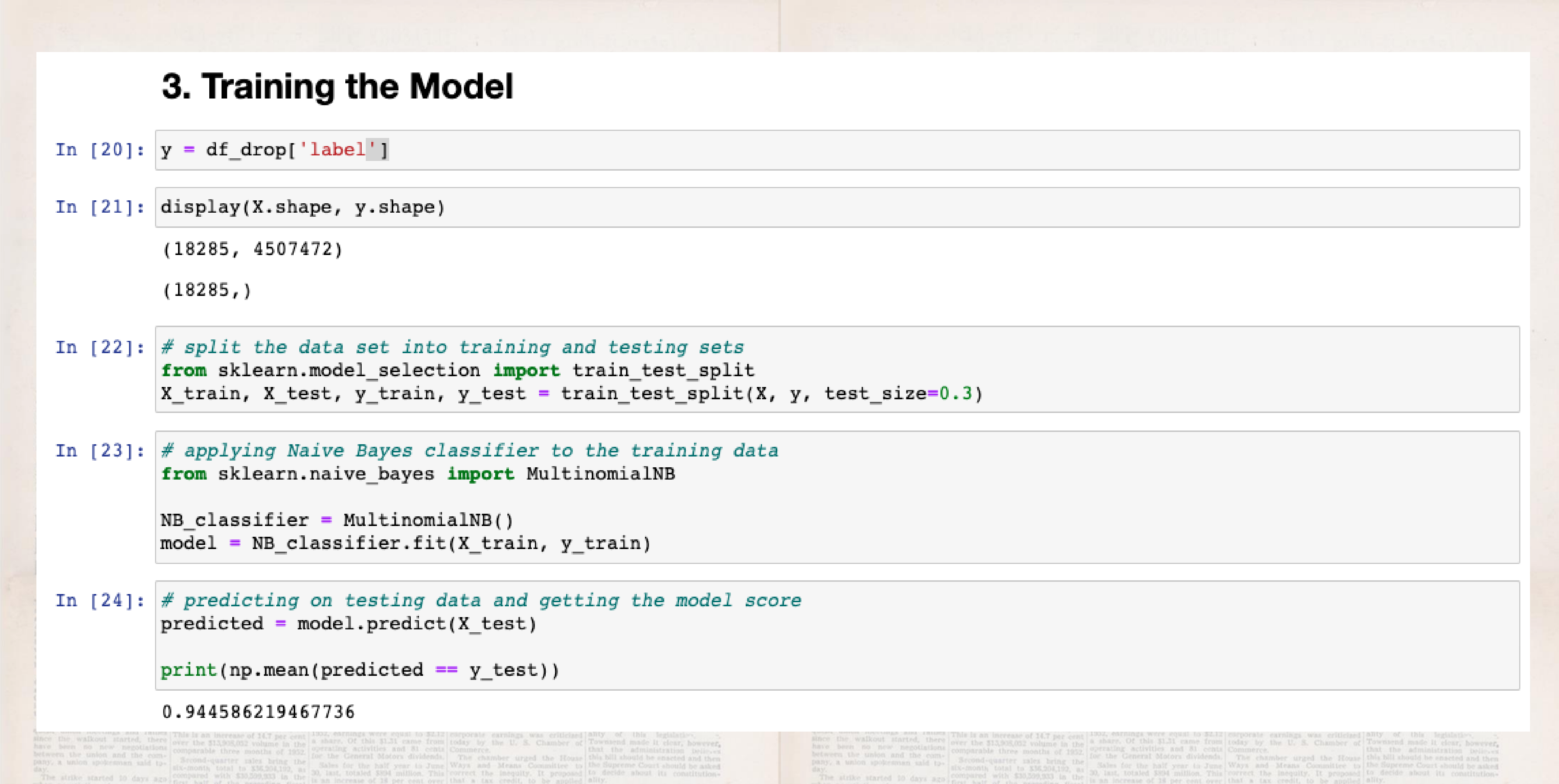
Preprocessing



```
In [23]: # removing stopwords (previously installed nltk in the PythonML env: python -m nltk.downloader
stop_words_removed = []
for i in range(0, len(df_clean['all'])):
    stop_words = [w for w in df_clean['all'][i] if w not in stopwords.words('english')]
    stop_words = ' '.join(stop_words)
    stop_words_removed.append(stop_words)
print(stop_words_removed[0])
```

```
house dem aide didnt even see comey letter jason chaffetz tweeted darrell lucus house dem aid
e didnt even see comey letter jason chaffetz tweeted darrell lucus october 30 2016 subscribe
jason chaffetz stump american fork utah image courtesy michael jolley available creative comm
onsby license apologies keith olbermann doubt worst person world weekfbi director james comey
according house democratic aide looks like also know secondworst person well turns comey sent
nowinfamous letter announcing fbi looking emails may related hillary clinton email server ran
king democrats relevant committees didnt hear comey found via tweet one republican committee
chairmen know comey notified republican chairmen democratic ranking members house intelligenc
e judiciary oversight committees agency reviewing emails recently discovered order see contai
ned classified information long letter went oversight committee chairman jason chaffetz set p
olitical world ablaze tweet fbi dir informed fbi learned existence emails appear pertinent in
vestigation case reopened jason chaffetz jasoninthehouse october 28 2016 course know case com
ey actually saying reviewing emails light unrelated casewhich know anthony weiner sexting tee
nager apparently little things facts didnt matter chaffetz utah republican already vowed init
iate raft investigations hillary winsat least two years worth possibly entire term worth appa
rently chaffetz thought fbi already work himresulting tweet briefly roiled nation cooler head
s realized dud according senior house democratic aide misreading letter may least chaffetz si
ns aide told shareblue boss democrats didnt even know comey letter timeand found checked twit
ter democratic ranking members relevant committees didnt receive comey letter republican chai
rmen fact democratic ranking members receive chairman oversight government reform committee j
ason chaffetz tweeted made public let see weve got right fbi director tells chaffetz gop comm
ittee chairmen major development potentially politically explosive investigation neither chaf
fetz colleagues courtesy let democratic counterparts know instead according aide made find tw
itter already talk daily kos comey provided advance notice letter chaffetz republicans giving
time turn spin machine may make good theater nothing far even suggests case nothing far sugge
sts comey anything grossly incompetent tonedeaf suggest however chaffetz acting way makes dan
burton darrell issa look like models responsibility bipartisanship didnt even decency notify
ranking member elijah cummings something explosive doesnt trample basic standards fairness do
nt know granted likely chaffetz answer sits ridiculously republican district anchored provo o
rem cook partisan voting index r25 gave mitt romney punishing 78 percent vote 2012 moreover r
epublican house leadership given full support chaffetz planned fishing expedition doesnt mean
cant turn hot lights textbook example house become republican control also second worst perso
n world darrell lucus darrell 30something graduate university north carolina considers journa
```

Training the Model



3. Training the Model

```
In [20]: y = df_drop['label']
```

```
In [21]: display(X.shape, y.shape)
```

```
(18285, 4507472)

(18285,)
```

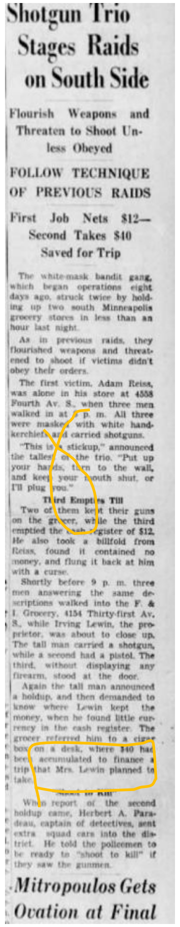
```
In [22]: # split the data set into training and testing sets
         from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

```
In [23]: # applying Naive Bayes classifier to the training data
         from sklearn.naive_bayes import MultinomialNB

         NB_classifier = MultinomialNB()
         model = NB_classifier.fit(X_train, y_train)
```

```
In [24]: # predicting on testing data and getting the model score
         predicted = model.predict(X_test)

         print(np.mean(predicted == y_test))
```

```
0.944586219467736
```

Naïve Bayes Classifier



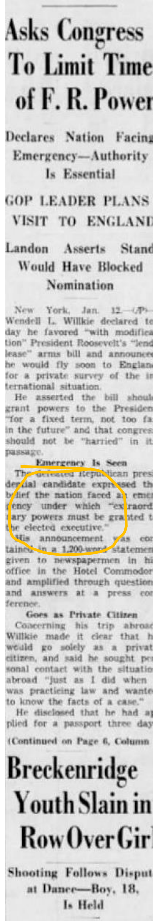
Shotgun Trio Stages Raids on South Side

Flourish Weapons and Threaten to Shoot Unless Obeyed

FOLLOW TECHNIQUE OF PREVIOUS RAIDS

First Job Nets $12—Second Takes $10 Saved for Trip

Mitropoulos Gets Ovation at Final



Asks Congress To Limit Time of F. R. Power

Declares Nation Facing Emergency—Authority Is Essential

GOP LEADER PLANS VISIT TO ENGLAND

Landon Asserts Stand Would Have Blocked Nomination

Emergency Is Seen

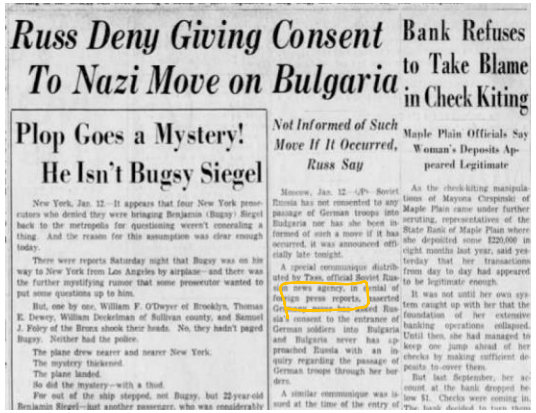
Goes as Private Citizen

Breckenridge Youth Slain in Row Over Gir

Shooting Follows Disput at Dance—Boy, 18, Is Held

Naive bayes combines both probability & bayes theorem to predict the outcome of a text, then categorizes it to a tag word.

In simpler terms, tag words are like "categories", we are trying to decipher snippets of text to be put into these categories.



Russ Deny Giving Consent To Nazi Move on Bulgaria

Plop Goes a Mystery! He Isn't Bugsy Siegel

Not Informed of Such Move If It Occurred, Russ Say

Bank Refuses to Take Blame in Check Kiting

Maple Plain Officials Say Woman's Deposits Appeared Legitimate

Critics Answer Shipstead View on War Policy

'U' Dean Lashes at Senator —'Confused' and 'Honest Statement' Others Say

A good example of naive bayes is a primary email box versus a spam email box. Our best outcome was the combination of 3 fields: title, author, and text.

DANGER OFF FLORIDA! IT MAY EXPLODE!

Minneapolis Morning Tribune

Masked Gang Robs 2 More Shops

CLASSIFIEDS

BUSINESS

REAL ESTATE