PROJECT: 8

PROJECT ID: Proj_225023_Team_3

NAME: S Paul Benjamin Felix

FAKE NEWS DETECTION USING NLP

PHASE 3 – DEVELOPMENT PART 1

Begin building your project by loading and preprocessing the dataset.

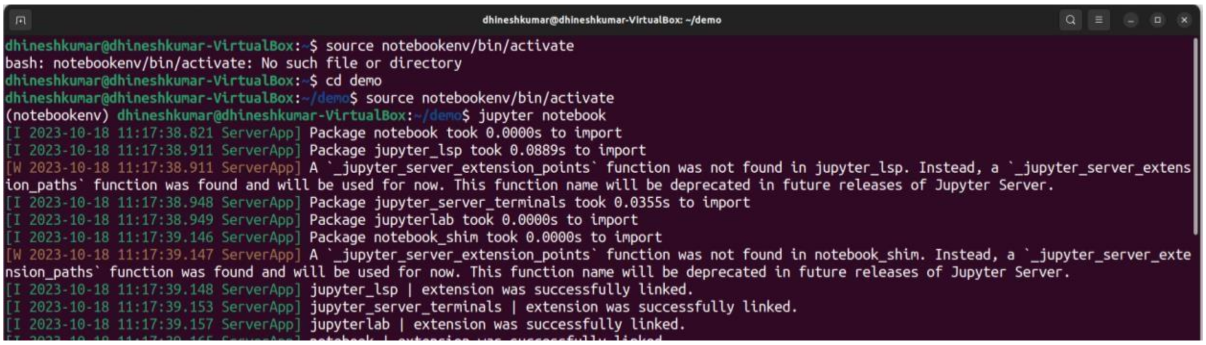
Dataset link: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

PRE-REQUISITES

1. Datasets **Fake.csv** and **True.csv** from Kaggle
2. Jupyter Notebook

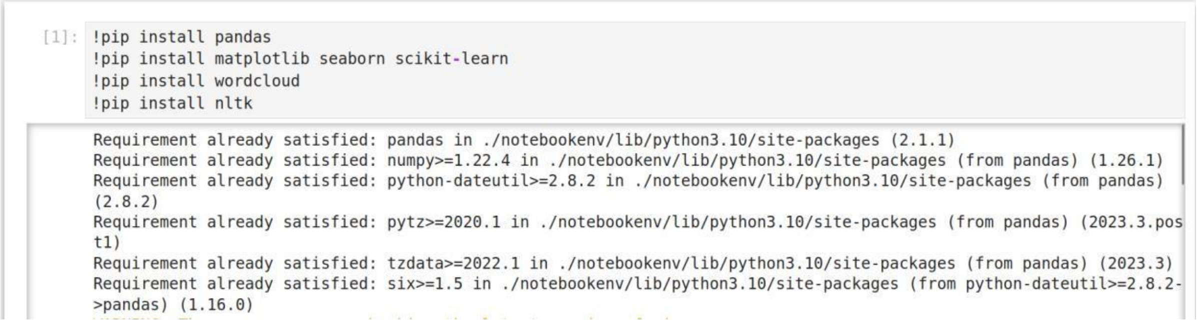
CODING

Step 1: Start Jupyter Notebook in any directory of choice.



```
dhineshkumar@dhineshkumar-VirtualBox:~$ source notebookenv/bin/activate
bash: notebookenv/bin/activate: No such file or directory
dhineshkumar@dhineshkumar-VirtualBox:~$ cd demo
dhineshkumar@dhineshkumar-VirtualBox:~/demo$ source notebookenv/bin/activate
(notebookenv) dhineshkumar@dhineshkumar-VirtualBox:~/demo$ jupyter notebook
[I 2023-10-18 11:17:38.821 ServerApp] Package notebook took 0.0000s to import
[I 2023-10-18 11:17:38.911 ServerApp] Package jupyter_lsp took 0.0889s to import
[W 2023-10-18 11:17:38.911 ServerApp] A `_jupyter_server_extension_points` function was not found in jupyter_lsp. Instead, a `_jupyter_server_extens
ion_paths` function was found and will be used for now. This function name will be deprecated in future releases of Jupyter Server.
[I 2023-10-18 11:17:38.948 ServerApp] Package jupyter_server_terminals took 0.0355s to import
[I 2023-10-18 11:17:38.949 ServerApp] Package jupyterlab took 0.0000s to import
[I 2023-10-18 11:17:39.146 ServerApp] Package notebook_shim took 0.0000s to import
[W 2023-10-18 11:17:39.147 ServerApp] A `_jupyter_server_extension_points` function was not found in notebook_shim. Instead, a `_jupyter_server_exte
nsion_paths` function was found and will be used for now. This function name will be deprecated in future releases of Jupyter Server.
[I 2023-10-18 11:17:39.148 ServerApp] jupyter_lsp | extension was successfully linked.
[I 2023-10-18 11:17:39.153 ServerApp] jupyter_server_terminals | extension was successfully linked.
[I 2023-10-18 11:17:39.157 ServerApp] jupyterlab | extension was successfully linked.
```

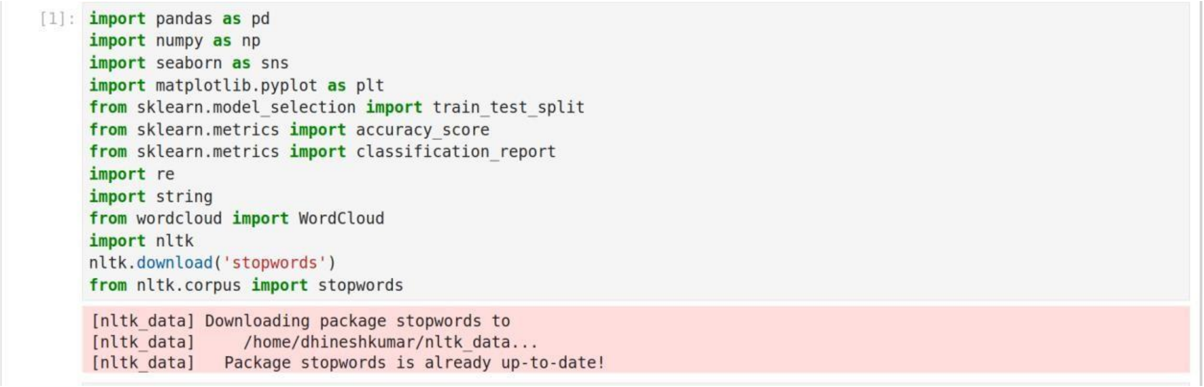
Step 2: Download required libraries in a new Notebook.



```
[1]: !pip install pandas
     !pip install matplotlib seaborn scikit-learn
     !pip install wordcloud
     !pip install nltk
```

```
Requirement already satisfied: pandas in ./notebookenv/lib/python3.10/site-packages (2.1.1)
Requirement already satisfied: numpy>=1.22.4 in ./notebookenv/lib/python3.10/site-packages (from pandas) (1.26.1)
Requirement already satisfied: python-dateutil>=2.8.2 in ./notebookenv/lib/python3.10/site-packages (from pandas)
(2.8.2)
Requirement already satisfied: pytz>=2020.1 in ./notebookenv/lib/python3.10/site-packages (from pandas) (2023.3.pos
t1)
Requirement already satisfied: tzdata>=2022.1 in ./notebookenv/lib/python3.10/site-packages (from pandas) (2023.3)
Requirement already satisfied: six>=1.5 in ./notebookenv/lib/python3.10/site-packages (from python-dateutil>=2.8.2-
>pandas) (1.16.0)
```

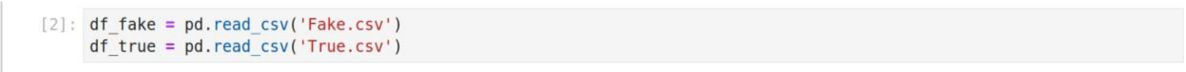
Step 3: Import the downloaded libraries according to the requirements of the project.



```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
import re
import string
from wordcloud import WordCloud
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
```

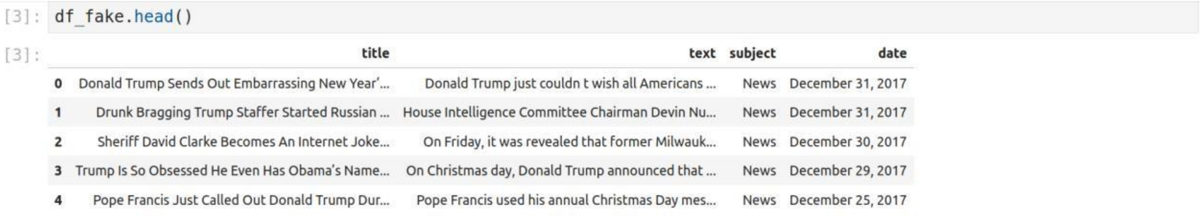
```
[nltk_data] Downloading package stopwords to
[nltk_data]     /home/dhineshkumar/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Step 4: Load the datasets using “read_csv” method to analyse and manipulate the data.



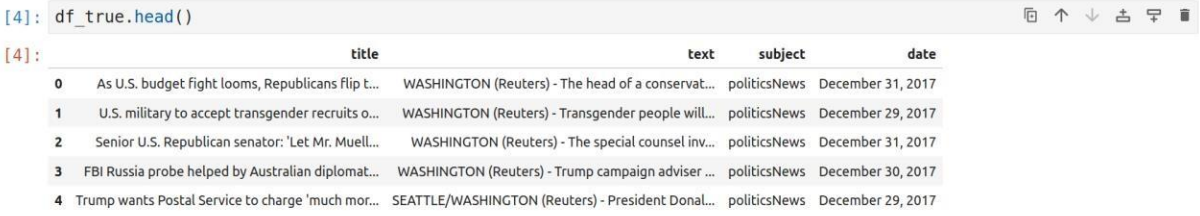
```
df_fake = pd.read_csv('Fake.csv')
df_true = pd.read_csv('True.csv')
```

Step 5: Using head( ) method to view the first 5 items in the dataset.



```
df_fake.head()
```

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year’...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama’s Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017



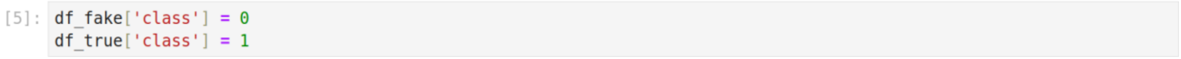
```
df_true.head()
```

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017

Step 6: Add a binary variable under class column to identify true from fake news and vice versa.

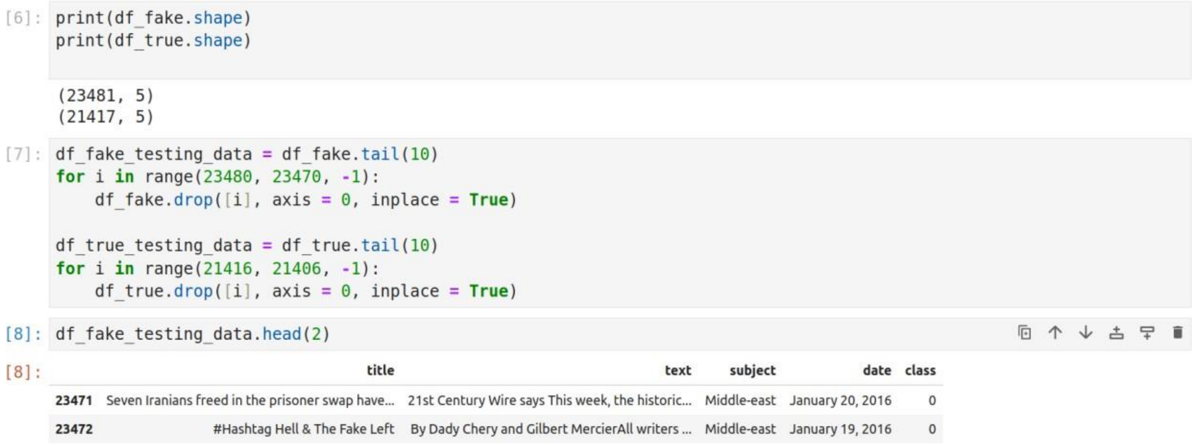
Fake news - 0

True news – 1



```
df_fake['class'] = 0
df_true['class'] = 1
```

Step 7: Separate some data from the dataset for manual testing that can be done at the last to manually check the accuracy of the machine learning model that has been built.



```
[6]: print(df_fake.shape)
     print(df_true.shape)

     (23481, 5)
     (21417, 5)
```

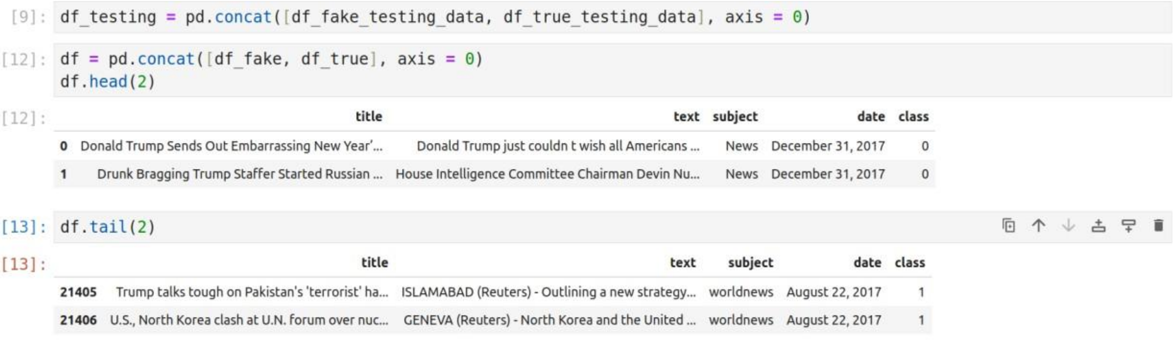
```
[7]: df_fake_testing_data = df_fake.tail(10)
     for i in range(23480, 23470, -1):
         df_fake.drop([i], axis = 0, inplace = True)

     df_true_testing_data = df_true.tail(10)
     for i in range(21416, 21406, -1):
         df_true.drop([i], axis = 0, inplace = True)
```

```
[8]: df_fake_testing_data.head(2)
```

[8]:

	title	text	subject	date	class
23471	Seven Iranians freed in the prisoner swap have...	21st Century Wire says This week, the historic...	Middle-east	January 20, 2016	0
23472	#Hashtag Hell & The Fake Left	By Dady Chery and Gilbert MercierAll writers ...	Middle-east	January 19, 2016	0

Step 8: Concatenate both the manual testing datasets into “df_testing” variable. Concatenate the other fake and true dataset into a variable “df”.



```
[9]: df_testing = pd.concat([df_fake_testing_data, df_true_testing_data], axis = 0)
```

```
[12]: df = pd.concat([df_fake, df_true], axis = 0)
      df.head(2)
```

[12]:

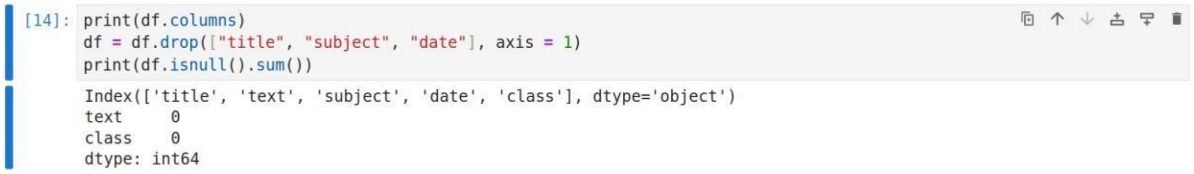
	title	text	subject	date	class
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0

```
[13]: df.tail(2)
```

[13]:

	title	text	subject	date	class
21405	Trump talks tough on Pakistan's 'terrorist' ha...	ISLAMABAD (Reuters) - Outlining a new strategy...	worldnews	August 22, 2017	1
21406	U.S., North Korea clash at U.N. forum over nuc...	GENEVA (Reuters) - North Korea and the United ...	worldnews	August 22, 2017	1

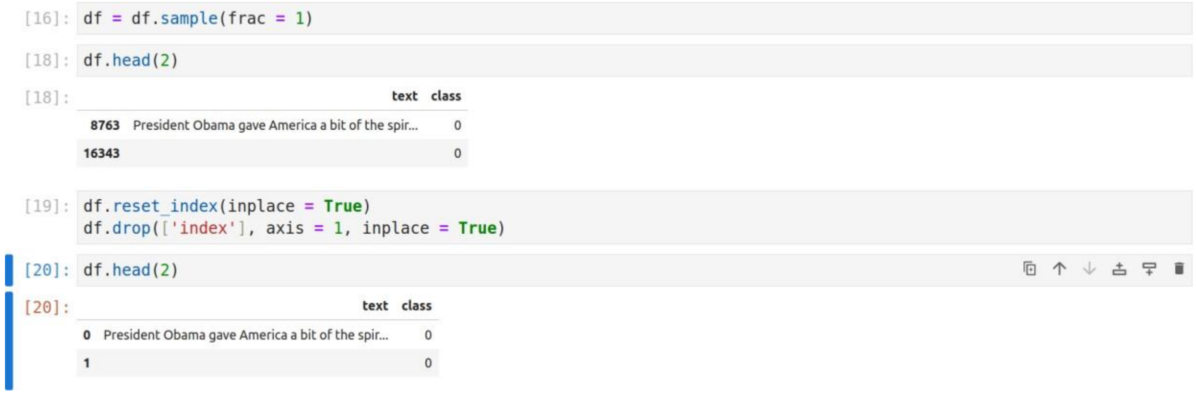
Step 9: Removing the unnecessary columns from the dataset to make the model more precise in classifying fake news from true news.



```
[14]: print(df.columns)
      df = df.drop(["title", "subject", "date"], axis = 1)
      print(df.isnull().sum())

      Index(['title', 'text', 'subject', 'date', 'class'], dtype='object')
      text     0
      class    0
      dtype: int64
```

“text” and “class” are the only columns required for building machine learning models. There are no null values in the dataset based on the output shown above.

Step 10: Shuffling the dataset and resetting the index values to prevent overfitting. Removing the additional index form the dataset.



```
[16]: df = df.sample(frac = 1)

[18]: df.head(2)
```

[18]:

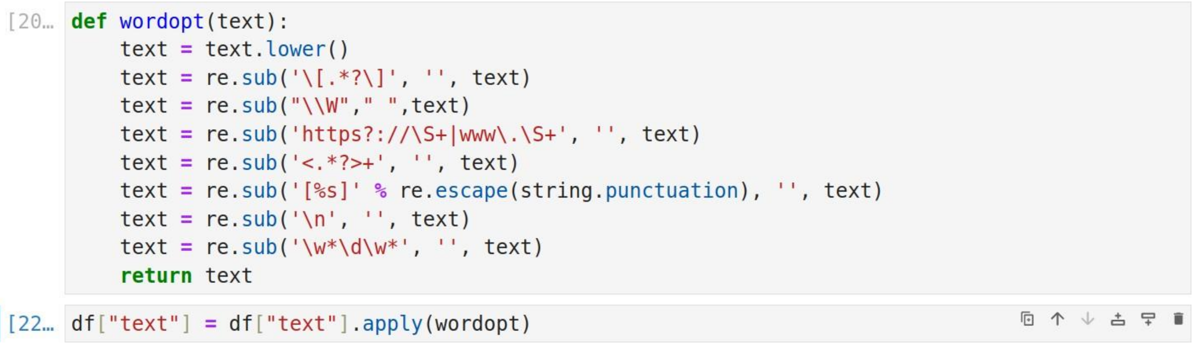
	text	class
8763	President Obama gave America a bit of the spir...	0
16343		0

```
[19]: df.reset_index(inplace = True)
      df.drop(['index'], axis = 1, inplace = True)

[20]: df.head(2)
```

[20]:

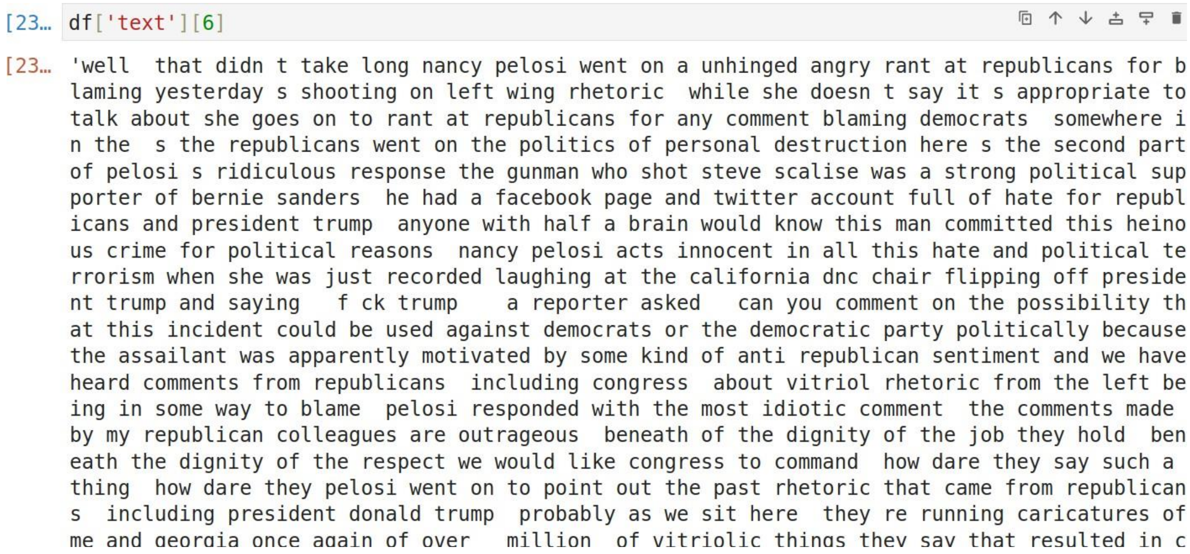
	text	class
0	President Obama gave America a bit of the spir...	0
1		0

Step 11: Change all the text into lowercase. Use regular expression to remove empty spaces, punctuations, html tags and escape characters. Apply the same function to all the text in the dataset.



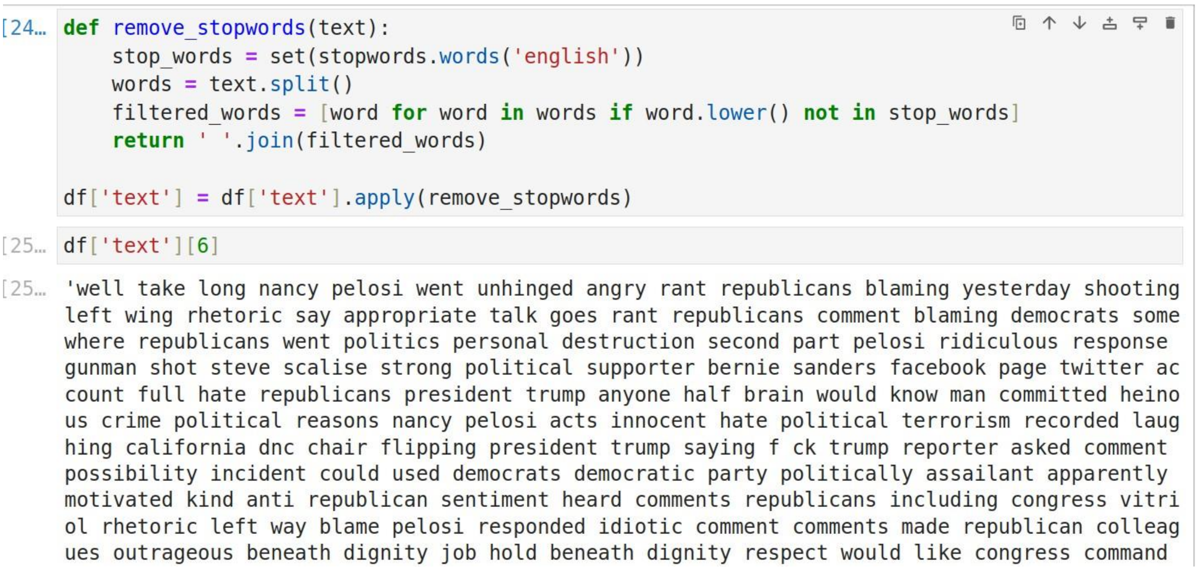
```
[20… def wordopt(text):
         text = text.lower()
         text = re.sub('\[.*?\]', '', text)
         text = re.sub("\\W"," ",text)
         text = re.sub('https?://\S+|www\.\S+', '', text)
         text = re.sub('<.*?>+', '', text)
         text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
         text = re.sub('\n', '', text)
         text = re.sub('\w*\d\w*', '', text)
         return text

[22… df["text"] = df["text"].apply(wordopt)
```

Print the text of a random index to view the affected text.



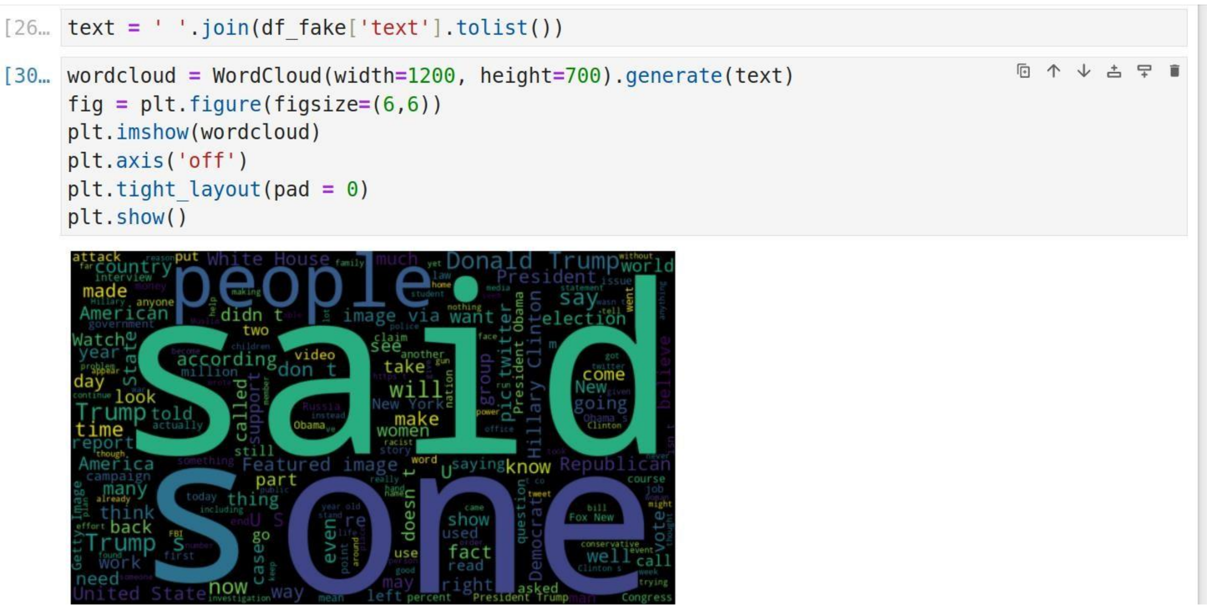
```
[23… df['text'][6]

[23… 'well  that didn t take long nancy pelosi went on a unhinged angry rant at republicans for b
     laming yesterday s shooting on left wing rhetoric  while she doesn t say it s appropriate to
     talk about she goes on to rant at republicans for any comment blaming democrats  somewhere i
     n the  s the republicans went on the politics of personal destruction here s the second part
     of pelosi s ridiculous response the gunman who shot steve scalise was a strong political sup
     porter of bernie sanders  he had a facebook page and twitter account full of hate for republ
     icans and president trump  anyone with half a brain would know this man committed this heino
     us crime for political reasons  nancy pelosi acts innocent in all this hate and political te
     rrorism when she was just recorded laughing at the california dnc chair flipping off preside
     nt trump and saying   f ck trump    a reporter asked   can you comment on the possibility th
     at this incident could be used against democrats or the democratic party politically because
     the assailant was apparently motivated by some kind of anti republican sentiment and we have
     heard comments from republicans  including congress  about vitriol rhetoric from the left be
     ing in some way to blame  pelosi responded with the most idiotic comment  the comments made
     by my republican colleagues are outrageous  beneath of the dignity of the job they hold  ben
     eath the dignity of the respect we would like congress to command  how dare they say such a
     thing  how dare they pelosi went on to point out the past rhetoric that came from republican
     s  including president donald trump  probably as we sit here  they re running caricatures of
     me and georgia once again of over   million  of vitriolic things they say that resulted in c
```

Step 12: Remove stop words from the text column using nltk library. It will eliminate articles and other words that are not required for training a model.



```
[24… def remove_stopwords(text):
         stop_words = set(stopwords.words('english'))
         words = text.split()
         filtered_words = [word for word in words if word.lower() not in stop_words]
         return ' '.join(filtered_words)

     df['text'] = df['text'].apply(remove_stopwords)

[25… df['text'][6]

[25… 'well take long nancy pelosi went unhinged angry rant republicans blaming yesterday shooting
     left wing rhetoric say appropriate talk goes rant republicans comment blaming democrats some
     where republicans went politics personal destruction second part pelosi ridiculous response
     gunman shot steve scalise strong political supporter bernie sanders facebook page twitter ac
     count full hate republicans president trump anyone half brain would know man committed heino
     us crime political reasons nancy pelosi acts innocent hate political terrorism recorded laug
     hing california dnc chair flipping president trump saying f ck trump reporter asked comment
     possibility incident could used democrats democratic party politically assailant apparently
     motivated kind anti republican sentiment heard comments republicans including congress vitri
     ol rhetoric left way blame pelosi responded idiotic comment comments made republican colleag
     ues outrageous beneath dignity job hold beneath dignity respect would like congress command
```

Step 13: (OPTIONAL) To view the frequency of the words in the fake dataset, it is preferrable to use word cloud. The same can be applied to True dataset as well.



```
[26… text = ' '.join(df_fake['text'].tolist())

[30… wordcloud = WordCloud(width=1200, height=700).generate(text)
     fig = plt.figure(figsize=(6,6))
     plt.imshow(wordcloud)
     plt.axis('off')
     plt.tight_layout(pad = 0)
     plt.show()
```

Data importing and preprocessing steps for the Fake News Detection using NLP project has been completed.

In the next phase, feature engineering and model training will be carried out to complete building the model.