In [15]:

```python
import pandas
import numpy as np
import matplotlib.pyplot as plt

#Seaborn is an open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data
#analysis. Seaborn works easily with dataframes and the Pandas library. The graphs created can also be customized easily.
import seaborn as sns
from sklearn.feature_extraction.text import CountVectorizer
#used for bag of words and it extract feature from text document
from sklearn.feature_extraction.text import TfidfTransformer
#tfid = term frequency is used for
from sklearn import feature_extraction, linear_model, model_selection, preprocessing
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
#cross velidation when setting different parameters
```

In [16]:

```python
fake = pandas.read_csv(r"C:\Users\iisha\Downloads\Fake.csv")
true = pandas.read_csv(r"C:\Users\iisha\Downloads\True.csv")
```

In [17]:

```python
fake.shape
```

Out[17]: (23481, 4)

In [18]:

```python
true.shape
```

Out[18]: (21417, 4)

In [19]:

```python
fake.head()
```

Out[19]:

	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

In [20]:

```python
true.head()
```

Out[20]:

	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

In [21]:

```python
# add flag to track fake and real
fake['target'] = 'fake'
true['target'] = 'true'
```

In [22]:

```python
fake.head()
```

Out[22]:

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

In [23]:

```python
true.head()
```

Out[23]:

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

In [24]:

```python
# combine datasets
data = pandas.concat([fake, true]).reset_index(drop = True)
data.shape
```

Out[24]: (44898, 5)

In [25]:

```python
data.head()
```

Out[25]:

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

In [26]:

```python
data.tail()
```

Out[26]:

	title	text	subject	date	target
44893	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	true
44894	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	true
44895	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	true
44896	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	true
44897	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	true

In [27]:

```python
# shuffle the data

from sklearn.utils import shuffle
data = shuffle(data)
data = data.reset_index(drop=True)
```

In [28]:

```python
data.head()
```

Out[28]:

	title	text	subject	date	target
0	IRRATIONAL GEORGETOWN PROFESSOR Has Month-Long...	A Georgetown University associate professor ha...	left-news	Dec 28, 2016	fake
1	U.S. lawmakers want documents on Russia electi...	WASHINGTON (Reuters) - The House of Representa...	politicsNews	January 25, 2017	true
2	NSA chief, intelligence director won't comment...	WASHINGTON (Reuters) - Two top U.S. intelligen...	politicsNews	June 7, 2017	true
3	Obama's Party With #BlackLivesMatter Organizer...	The Obama s are classless we knew that. The Ob...	left-news	Dec 5, 2015	fake
4	OBAMA'S EPA GESTAPO TO SKIP Hearing On CO Mine...	Priorities priorities Arizona Sen. John McCain...	left-news	Apr 7, 2016	fake

In [29]:

```python
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 44898 entries, 0 to 44897
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   title    44898 non-null  object
 1   text     44898 non-null  object
 2   subject  44898 non-null  object
 3   date     44898 non-null  object
 4   target   44898 non-null  object
dtypes: object(5)
memory usage: 1.7+ MB
```

In [30]:

```python
# removing date
data.drop(['date'],axis=1,inplace=True)
data.head()
```

Out[30]:

	title	text	subject	target
0	IRRATIONAL GEORGETOWN PROFESSOR Has Month-Long...	A Georgetown University associate professor ha...	left-news	fake
1	U.S. lawmakers want documents on Russia electi...	WASHINGTON (Reuters) - The House of Representa...	politicsNews	true
2	NSA chief, intelligence director won't comment...	WASHINGTON (Reuters) - Two top U.S. intelligen...	politicsNews	true
3	Obama's Party With #BlackLivesMatter Organizer...	The Obama s are classless we knew that. The Ob...	left-news	fake
4	OBAMA'S EPA GESTAPO TO SKIP Hearing On CO Mine...	Priorities priorities Arizona Sen. John McCain...	left-news	fake

In [31]:

```python
data.drop(['title'],axis=1,inplace=True)
data.head()
```

Out[31]:

	text	subject	target
0	A Georgetown University associate professor ha...	left-news	fake
1	WASHINGTON (Reuters) - The House of Representa...	politicsNews	true
2	WASHINGTON (Reuters) - Two top U.S. intelligen...	politicsNews	true
3	The Obama s are classless we knew that. The Ob...	left-news	fake
4	Priorities priorities Arizona Sen. John McCain...	left-news	fake

In [32]:

```python
#convert lower case
data['text'] = data['text'].apply(lambda x: x.lower())
data.head()
```

Out[32]:

	text	subject	target
0	a georgetown university associate professor ha...	left-news	fake
1	washington (reuters) - the house of representa...	politicsNews	true
2	washington (reuters) - two top u.s. intelligen...	politicsNews	true
3	the obama s are classless we knew that. the ob...	left-news	fake
4	priorities priorities arizona sen. john mccain...	left-news	fake

In [33]:

```python
# remove punctuation

import string

def punctuation_removal(text):
    all_list = [char for char in text if char not in string.punctuation]
    clean_str = ''.join(all_list)
    return clean_str

data['text'] = data['text'].apply(punctuation_removal)
```

In [34]:

```python
data.head()
```

Out[34]:

	text	subject	target
0	a georgetown university associate professor ha...	left-news	fake
1	washington reuters the house of representativ...	politicsNews	true
2	washington reuters two top us intelligence of...	politicsNews	true
3	the obama s are classless we knew that the oba...	left-news	fake
4	priorities priorities arizona sen john mccain ...	left-news	fake

In [35]:

```python
# Removing stopwords
##NLTK is a standard python library with prebuilt functions and utilities for the ease of use and implementation.
##It is one of the most used libraries for natural language processing and computational linguistics.

import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
stop = stopwords.words('english')

data['text'] = data['text'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop)]))
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\iisha\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [36]:

```python
data.head()
```

Out[36]:

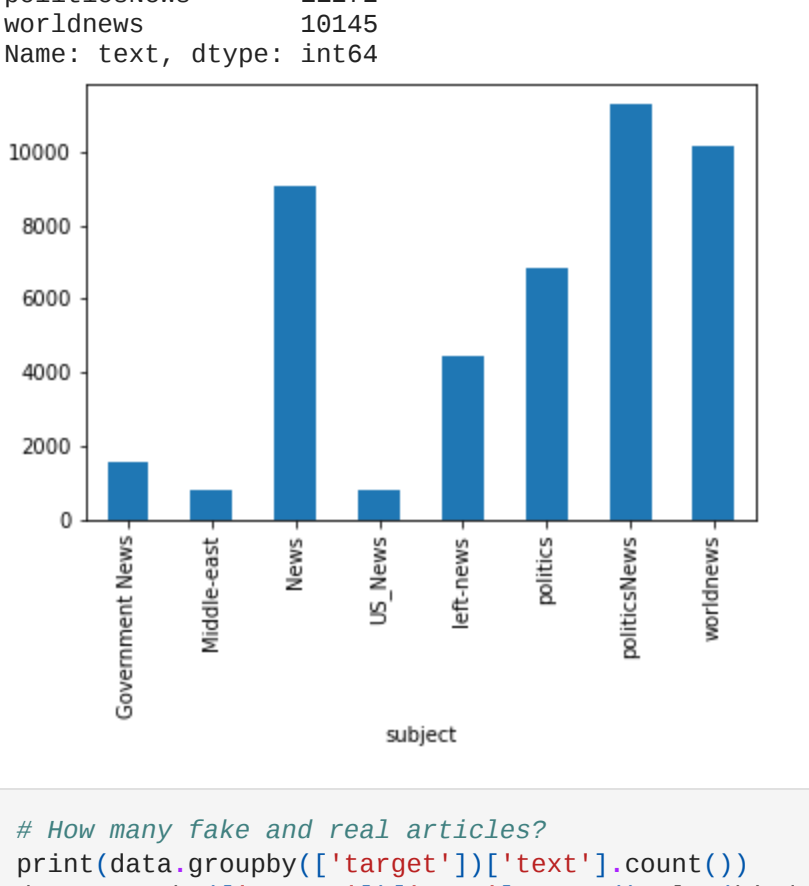
	text	subject	target
0	georgetown university associate professor mont...	left-news	fake
1	washington reuters house representatives intel...	politicsNews	true
2	washington reuters two top us intelligence off...	politicsNews	true
3	obama classless knew obama guest list holiday ...	left-news	fake
4	priorities priorities arizona sen john mccain ...	left-news	fake

Basic data exploration

In [37]:

```python
# How many articles per subject?
print(data.groupby(['subject'])['text'].count())
data.groupby(['subject'])['text'].count().plot(kind="bar")
plt.show()
```

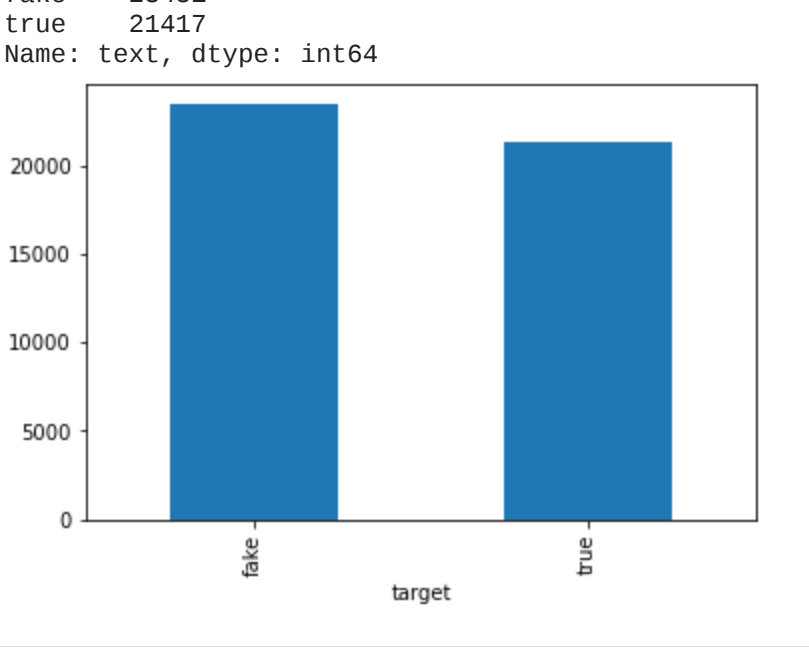
```
subject
Government News     1570
Middle-east          778
News                9050
US_News              783
left-news           4459
politics            6841
politicsNews       11272
worldnews          10145
Name: text, dtype: int64
```



In [38]:

```python
# How many fake and real articles?
print(data.groupby(['target'])['text'].count())
data.groupby(['target'])['text'].count().plot(kind="bar")
plt.show()
```

```
target
fake    23481
true    21417
Name: text, dtype: int64
```



In [39]:

```python
!pip install wordcloud
```

```
Requirement already satisfied: wordcloud in c:\users\iisha\anaconda3\lib\site-packages (1.8.1)
Requirement already satisfied: matplotlib in c:\users\iisha\anaconda3\lib\site-packages (from wordcloud) (3.3.4)
Requirement already satisfied: numpy>=1.6.1 in c:\users\iisha\anaconda3\lib\site-packages (from wordcloud) (1.20.1)
Requirement already satisfied: pillow in c:\users\iisha\anaconda3\lib\site-packages (from wordcloud) (8.2.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in c:\users\iisha\anaconda3\lib\site-packages (from matplotlib->wordcloud) (2.4.7)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\iisha\anaconda3\lib\site-packages (from matplotlib->wordcloud) (2.8.1)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\iisha\anaconda3\lib\site-packages (from matplotlib->wordcloud) (1.3.1)
Requirement already satisfied: cycler>=0.10 in c:\users\iisha\anaconda3\lib\site-packages (from matplotlib->wordcloud) (0.10.0)
Requirement already satisfied: six in c:\users\iisha\anaconda3\lib\site-packages (from cycler>=0.10->matplotlib->wordcloud) (1.15.0)
```

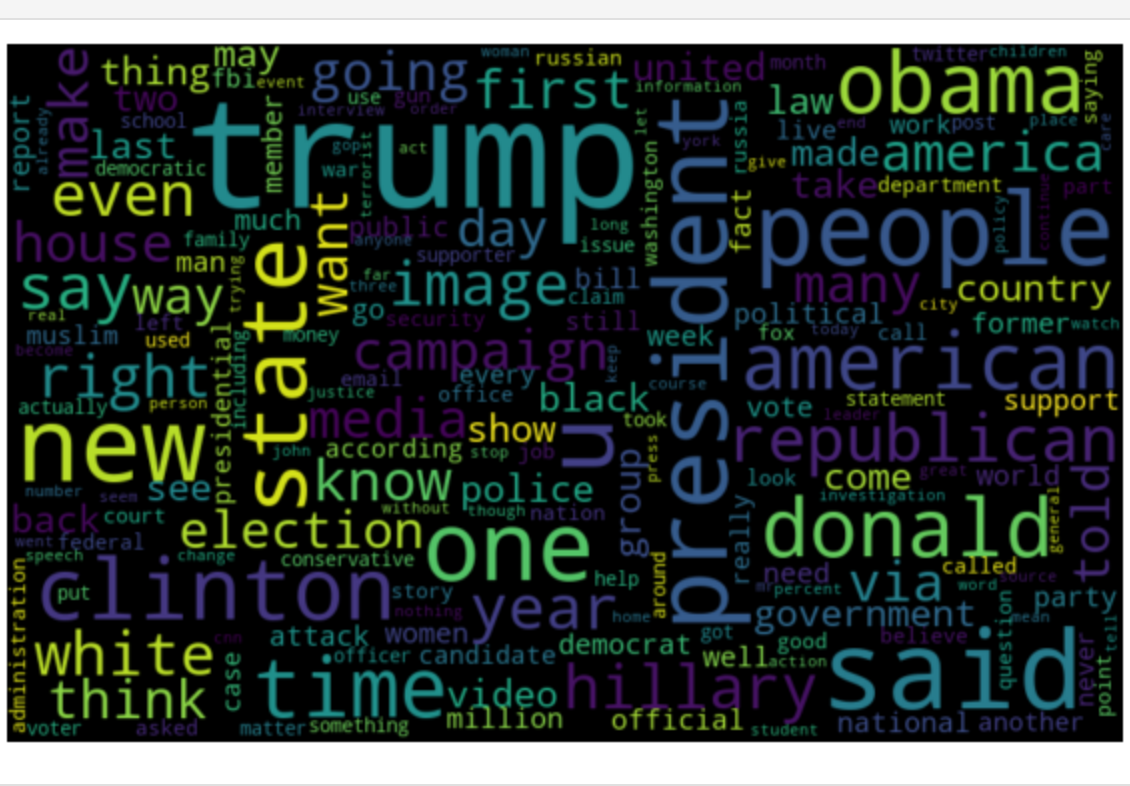
In [40]:

```python
# Word cloud for fake news
from wordcloud import WordCloud

fake_data = data[data["target"] == "fake"]
all_words = ' '.join([text for text in fake_data.text])

wordcloud = WordCloud(width= 800, height= 500,
                          max_font_size = 110,
                          collocations = False).generate(all_words)

plt.figure(figsize=(10,7))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```



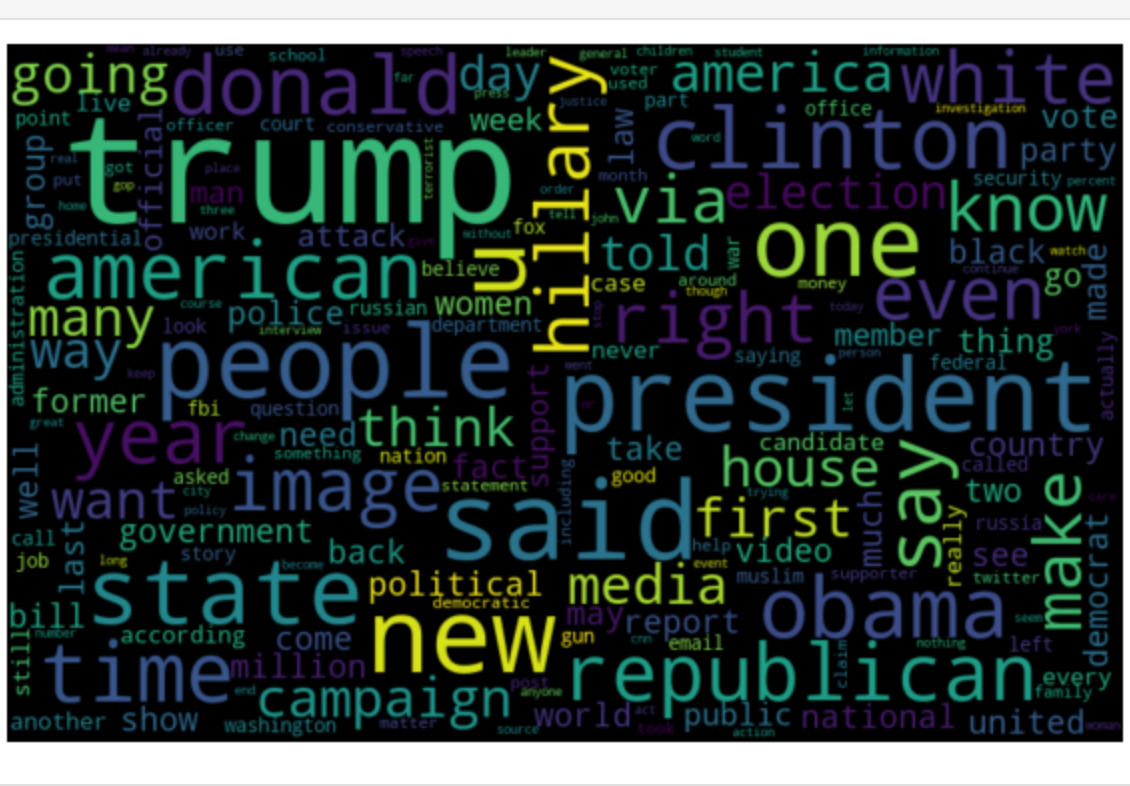
In [41]:

```python
# Word cloud for real news
from wordcloud import WordCloud

real_data = data[data["target"] == "true"]
all_words = ' '.join([text for text in fake_data.text])

wordcloud = WordCloud(width= 800, height= 500,
                          max_font_size = 110,
                          collocations = False).generate(all_words)

plt.figure(figsize=(10,7))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```



In [42]:

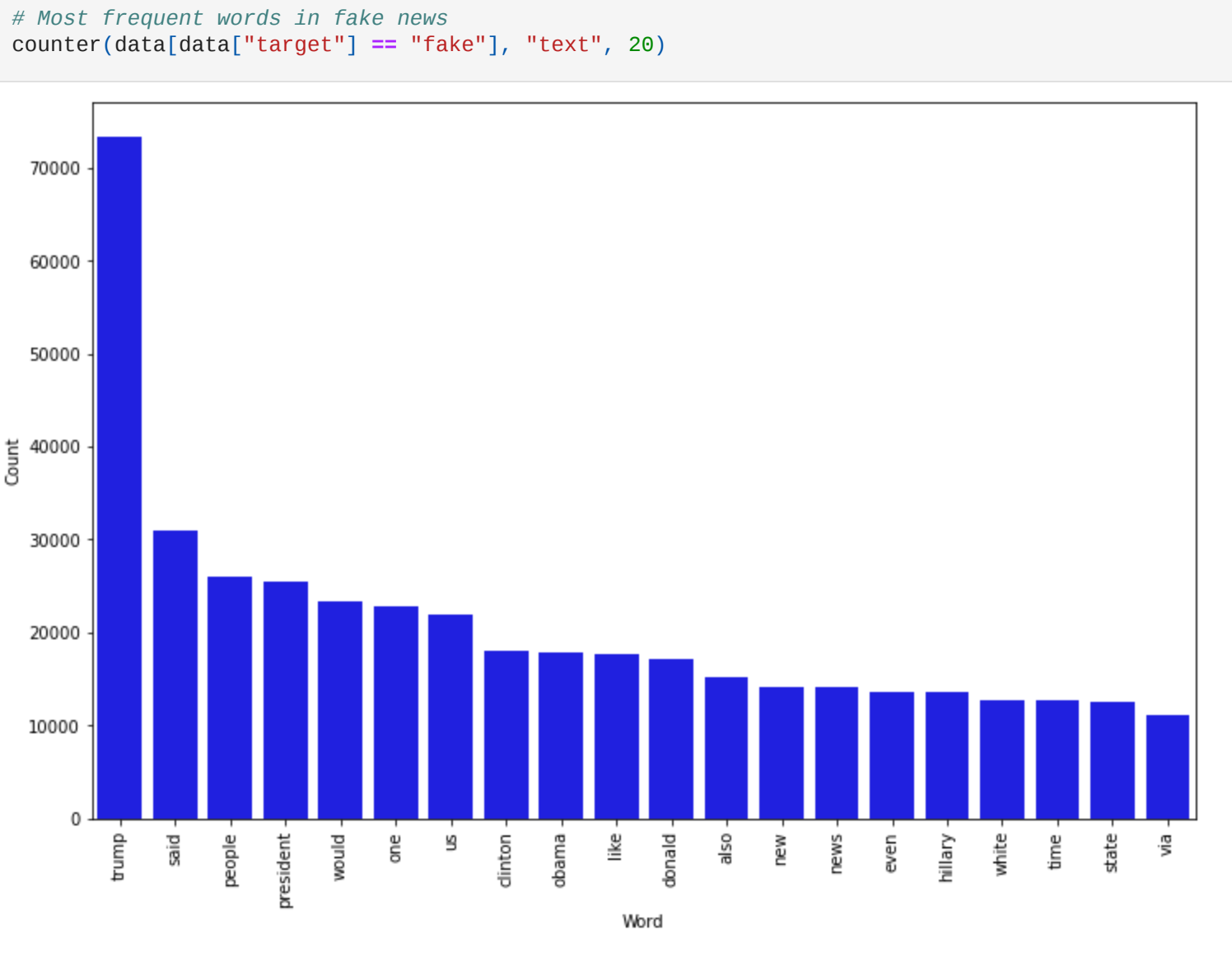
```python
# Most frequent words counter
from nltk import tokenize

token_space = tokenize.WhitespaceTokenizer()

def counter(text, column_text, quantity):
    all_words = ' '.join([text for text in text[column_text]])
    token_phrase = token_space.tokenize(all_words)
    frequency = nltk.FreqDist(token_phrase)
    df_frequency = pandas.DataFrame({"Word": list(frequency.keys()),
                                   "Frequency": list(frequency.values())})
    df_frequency = df_frequency.nlargest(columns = "Frequency", n = quantity)
    plt.figure(figsize=(12,8))
    ax = sns.barplot(data = df_frequency, x = "Word", y = "Frequency", color = 'blue')
    ax.set(ylabel = "Count")
    plt.xticks(rotation='vertical')
    plt.show()
```

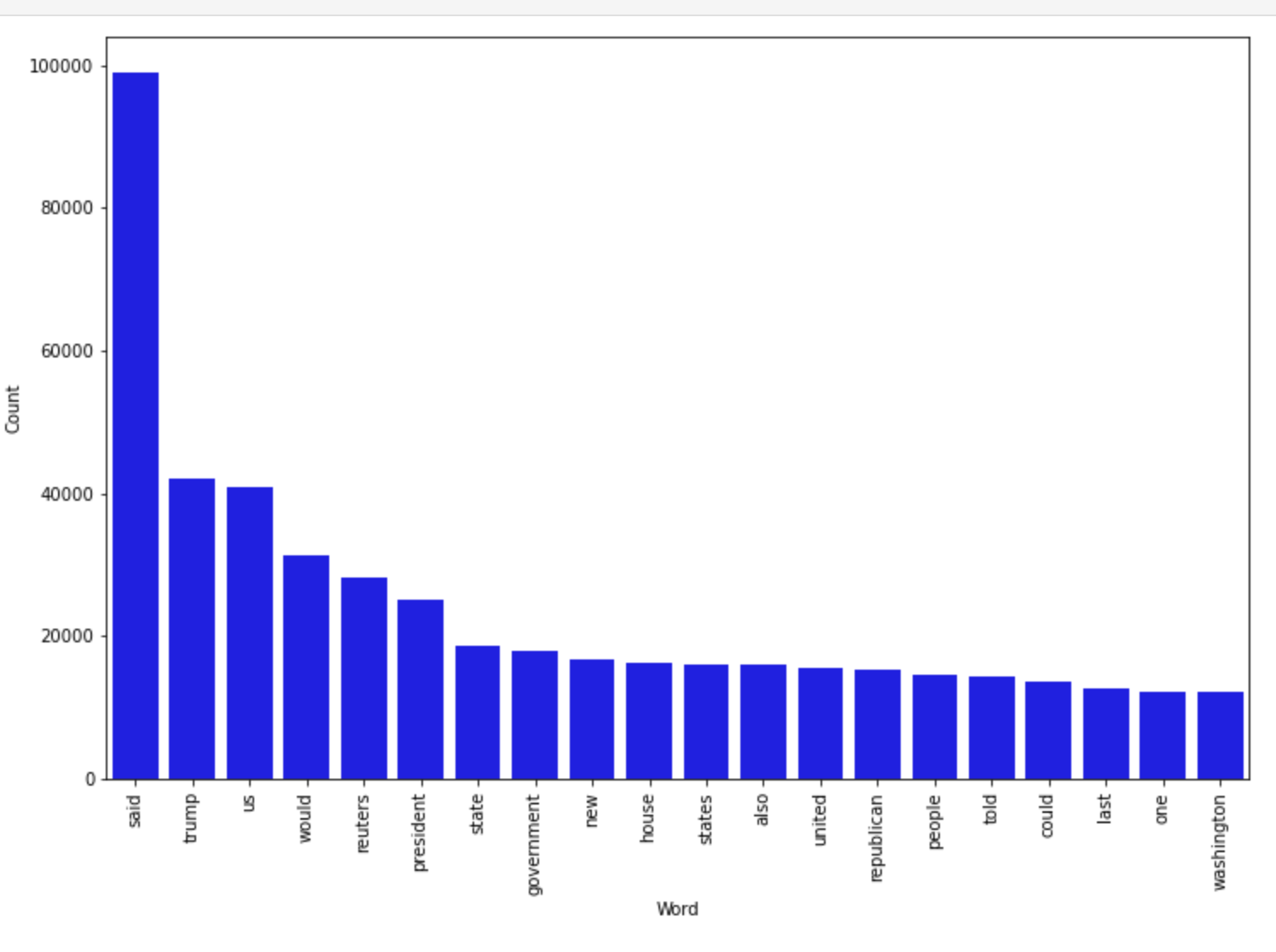
In [43]:

```python
# Most frequent words in fake news
counter(data[data["target"] == "fake"], "text", 20)
```



In [44]:

```python
# Most frequent words in real news
counter(data[data["target"] == "true"], "text", 20)
```



In [45]:

```python
# Function to plot the confusion matrix
from sklearn import metrics
import itertools

def plot_confusion_matrix(cm, classes,
                          normalize=False,
                          title='Confusion matrix',
                          cmap=plt.cm.Blues):

    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=45)
    plt.yticks(tick_marks, classes)

    if normalize:
        cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        print("Normalized confusion matrix")
    else:
        print('Confusion matrix, without normalization')

    thresh = cm.max() / 2.
    for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
        plt.text(j, i, cm[i, j],
                 horizontalalignment="center",
                 color="white" if cm[i, j] > thresh else "black")

    plt.tight_layout()
    plt.ylabel('True label')
    plt.xlabel('Predicted label')
```

Split data

In [46]:

```python
# Split the data
X_train,X_test,y_train,y_test = train_test_split(data['text'], data.target, test_size=0.2, random_state=42)
```

In [47]:

```python
X_train.head()
```

Out[47]:

```
36335    21st century wire says censorship running ramp...
12384    gothenburg sweden reuters deadlock divorce tal...
24419    washington reuters republican us senators john...
24740    new york reuters us republican presidential ca...
27039    dhaka reuters indonesia foreign minister fly b...
Name: text, dtype: object
```

In [48]:

```python
y_train.head()
```

Out[48]:

```
36335    fake
12384    true
24419    true
24740    true
27039    true
Name: target, dtype: object
```

Decision Tree Classifier

In [49]:

```python
from sklearn.tree import DecisionTreeClassifier

# Vectorizing and applying TF-IDF
pipe = Pipeline([('vect', CountVectorizer()),
                 ('tfidf', TfidfTransformer()),
                 ('model', DecisionTreeClassifier(criterion= 'entropy',
                                           max_depth = 20,
                                           splitter='best',
                                           random_state=42))])
# Fitting the model
model = pipe.fit(X_train, y_train)

# Accuracy
prediction = model.predict(X_test)
print("accuracy: {}%".format(round(accuracy_score(y_test, prediction)*100,2)))
```

accuracy: 99.61%

In [ ]: