D:\AI\tweets.csv

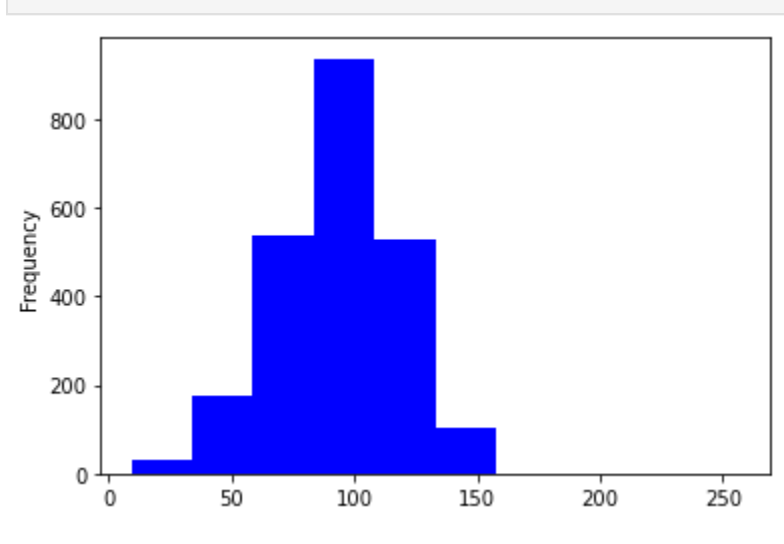
In [53]:
```python
import numpy as np
import pandas as pd
import re #text cleaning (preprocessing)
import nltk #natural language toolkit, used for preprocessing
import string
from nltk.stem.porter import PorterStemmer
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
```

In [54]:
```python
df_Whole = pd.read_csv("D:/AI/tweets.csv",nrows=2309)
df=df_Whole[["TWEET"]]
df_Whole.head()
```

Out[54]:

	TWEET_ID	TWEET	TWEET_Label
0	80080680482123700	seriously? racist mcdonaldâ€™s sign is obvious...	false
1	80084555733803000	hoax: mcdonald's issues official statement on ...	false
2	91728807081426900	#psa please do not drink any pepsi soda, a wor...	false
3	265953285247209000	deep-fried left wings demo-crab cakes barack-a...	false
4	273182568298450000	42 million dead in bloodiest black friday week...	false

In [55]:
```python
length_tweet = df['TWEET'].str.len().plot.hist(color = 'blue' , figsize = (6,4))
```



In [56]:
```python
df_Whole['length'] = df['TWEET'].str.len()
df_Whole['length'].describe()
```

Out[56]:
```
count    2308.000000
mean       92.633016
std        23.948201
min         9.000000
25%        77.000000
50%        95.000000
75%       110.000000
max       257.000000
Name: length, dtype: float64
```

In [57]:
```python
df_Whole[df_Whole['length']==100] ['TWEET'].iloc[0]
```

Out[57]:
```
"highly radioactive water is seeping into the ocean from japan's crippled fukushima nuclear plant URL"
```

In [58]:
```python
print(df_Whole)
```

```
                TWEET_ID                                              TWEET  \
0      80080680482123700  seriously? racist mcdonaldâ€™s sign is obvious...
1      80084555733803000  hoax: mcdonald's issues official statement on ...
2      91728807081426900  #psa please do not drink any pepsi soda, a wor...
3     265953285247209000  deep-fried left wings demo-crab cakes barack-a...
4     273182568298450000  42 million dead in bloodiest black friday week...
...                  ...                                                ...
2303  778689027918618000  black men may have cause to run from police, m...
2304  778949749156245000  a black man who runs from police merely might ...
2305  779633844680962000  this network of tunnels is from the stone age ...
2306  780436430732525000  well, this is a new one: chelsea clinton impli...
2307  780882510645370000  chelsea clinton implies here that marijuana ca...

     TWEET_Label  length
0          false      61
1          false      91
2          false     138
3          false     128
4          false      63
...          ...     ...
2303  unverified      78
2304  unverified     110
2305  unverified     117
2306  unverified     118
2307  unverified      65

[2308 rows x 4 columns]
```

In [59]:
```python
#.......................changing the casing of the words...............................
df = df_Whole.iloc[:,1:2]
df["TWEET"] = df['TWEET'].astype(str)
df["TWEET_lower"] = df["TWEET"].str.lower()
df.head()
```

Out[59]:

	TWEET	TWEET_lower
0	seriously? racist mcdonaldâ€™s sign is obvious...	seriously? racist mcdonaldâ€™s sign is obvious...
1	hoax: mcdonald's issues official statement on ...	hoax: mcdonald's issues official statement on ...
2	#psa please do not drink any pepsi soda, a wor...	#psa please do not drink any pepsi soda, a wor...
3	deep-fried left wings demo-crab cakes barack-a...	deep-fried left wings demo-crab cakes barack-a...
4	42 million dead in bloodiest black friday week...	42 million dead in bloodiest black friday week...

In [60]:
```python
#Remove punctuations
df.drop(["TWEET_lower"], axis=1, inplace=True)
puncremove = string.punctuation
def remove_punctuation(TWEET):
    return TWEET.translate(str.maketrans('','',puncremove))
df["TWEET_punct"] = df["TWEET"].apply(lambda TWEET: remove_punctuation(TWEET))
df.head()
```

Out[60]:

	TWEET	TWEET_punct
0	seriously? racist mcdonaldâ€™s sign is obvious...	seriously racist mcdonaldâ€™s sign is obviousl...
1	hoax: mcdonald's issues official statement on ...	hoax mcdonalds issues official statement on ra...
2	#psa please do not drink any pepsi soda, a wor...	psa please do not drink any pepsi soda a worke...
3	deep-fried left wings demo-crab cakes barack-a...	deepfried left wings democrab cakes barackamol...
4	42 million dead in bloodiest black friday week...	42 million dead in bloodiest black friday week...

In [61]:
```python
import numpy as np
import pandas as pd
import re #text cleaning (preprocessing)
import nltk #natural language toolkit, used for preprocessing
import string
from nltk.stem.porter import PorterStemmer
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
```

In [62]:
```python
stops = set(stopwords.words('english'))
print(stops)
```

```
{'those', 'mustn', 'in', 'no', 'or', "she's", 'me', 'down', 'most', 'had', "you're", 'it', 'which', "that'll", 'from', 'having', 'needn', 'wouldn', 'why', 'were', 'isn', 'more', 'w
on', 'up', 'am', 'just', 'ourselves', 'the', "mustn't", 'by', 'yourself', 'whom', 'did', 'and', 'what', 'after', 'haven', 're', 't', 'a', 'than', 'them', "you'll", 'o', 'hasn', "ha
ven't", 'is', 'was', 'between', 'doesn', 'about', 'once', 'm', 'does', 'for', 'ours', 'hers', 'into', "couldn't", 'don', "shan't", 'so', 'didn', 'been', 'out', 'myself', 'again',
'below', 'too', "should've", 'shouldn', "shouldn't", "won't", 'theirs', 'himself', 'him', 'herself', 'be', 'above', 'few', 'aren', "doesn't", 'your', "you've", 'do', 'all', 'with',
"hadn't", 'when', "it's", "you'd", 'there', 'being', 'mightn', 'doing', 'as', 'other', "wasn't", 'hadn', 'at', 'each', 'some', 'not', 'my', 'any', 'shan', 'i', 'you', "isn't", 'its
elf', 'these', 'here', 'y', 'then', 'yours', 'until', 'off', 'how', 'are', 'while', "mightn't", 'very', "needn't", "don't", 'over', 'of', 'under', 'themselves', 'd', 'its', 'becaus
e', 'll', 'before', 'their', 'her', 'own', 'further', "didn't", 'to', 'now', 'ma', 'an', 'will', 'against', 'where', 'couldn', 'during', 'his', 'nor', 'both', 'she', 've', 'only',
'he', 'ain', 'same', 'this', 'yourselves', 'on', 'such', 'should', 'wasn', 'through', 'they', "aren't", 'but', 'that', 's', "hasn't", "wouldn't", 'we', 'have', 'has', "weren't", 'i
f', 'weren', 'our', 'can', 'who'}
```

In [63]:
```python
STOPWORDS = set(stopwords.words('english'))
def remove_stopwords(TWEET):
    """custom function to remove the stopwords"""
    return " ".join([word for word in str(TWEET).split() if word not in STOPWORDS])

df["TWEET_stop"] = df["TWEET_punct"].apply(lambda TWEET: remove_stopwords(TWEET))
df.head()
```

Out[63]:

	TWEET	TWEET_punct	TWEET_stop
0	seriously? racist mcdonaldâ€™s sign is obvious...	seriously racist mcdonaldâ€™s sign is obviousl...	seriously racist mcdonaldâ€™s sign obviously h...
1	hoax: mcdonald's issues official statement on ...	hoax mcdonalds issues official statement on ra...	hoax mcdonalds issues official statement racis...
2	#psa please do not drink any pepsi soda, a wor...	psa please do not drink any pepsi soda a worke...	psa please drink pepsi soda worker company put...
3	deep-fried left wings demo-crab cakes barack-a...	deepfried left wings democrab cakes barackamol...	deepfried left wings democrab cakes barackamol...
4	42 million dead in bloodiest black friday week...	42 million dead in bloodiest black friday week...	42 million dead bloodiest black friday weekend...

In [64]:
```python
from collections import Counter
cnt = Counter()
for TWEET in df["TWEET_stop"].values:
    for word in TWEET.split():
        cnt[word] += 1

cnt.most_common(10)
```

Out[64]:
```
[('URL', 2990),
 ('police', 139),
 ('shot', 111),
 ('says', 102),
 ('ferguson', 92),
 ('breaking', 89),
 ('new', 86),
 ('us', 84),
 ('obama', 82),
 ('paul', 82)]
```

In [65]:
```python
freqwords = set([w for (w, wc) in cnt.most_common(10)])
def remove_freqwords(TWEET):
    """custom function to remove the frequent words"""
    return " ".join([word for word in str(TWEET).split() if word not in freqwords])

df["TWEET_stopfreq"] = df["TWEET_stop"].apply(lambda TWEET: remove_freqwords(TWEET))
df.head()
```

Out[65]:

	TWEET	TWEET_punct	TWEET_stop	TWEET_stopfreq
0	seriously? racist mcdonaldâ€™s sign is obvious...	seriously racist mcdonaldâ€™s sign is obviousl...	seriously racist mcdonaldâ€™s sign obviously h...	seriously racist mcdonaldâ€™s sign obviously hoax
1	hoax: mcdonald's issues official statement on ...	hoax mcdonalds issues official statement on ra...	hoax mcdonalds issues official statement racis...	hoax mcdonalds issues official statement racis...
2	#psa please do not drink any pepsi soda, a wor...	psa please do not drink any pepsi soda a worke...	psa please drink pepsi soda worker company put...	psa please drink pepsi soda worker company put...
3	deep-fried left wings demo-crab cakes barack-a...	deepfried left wings democrab cakes barackamol...	deepfried left wings democrab cakes barackamol...	deepfried left wings democrab cakes barackamol...
4	42 million dead in bloodiest black friday week...	42 million dead in bloodiest black friday week...	42 million dead bloodiest black friday weekend...	42 million dead bloodiest black friday weekend...

In [66]:
```python
df.drop(["TWEET_punct", "TWEET_stop"], axis=1, inplace=True)
```

In [67]:
```python
#stemming(playing-play)
stemmer = PorterStemmer()
def stem_words(TWEET):
    return " ".join([stemmer.stem(word) for word in TWEET.split()])

df["TWEET_stemmed"] = df["TWEET"].apply(lambda TWEET: stem_words(TWEET))
df.head(10)
```

Out[67]:

	TWEET	TWEET_stopfreq	TWEET_stemmed
0	seriously? racist mcdonaldâ€™s sign is obvious...	seriously racist mcdonaldâ€™s sign obviously hoax	seriously? racist mcdonaldâ€™ sign is obvious ...
1	hoax: mcdonald's issues official statement on ...	hoax mcdonalds issues official statement racis...	hoax: mcdonald' issu offici statement on racis...
2	#psa please do not drink any pepsi soda, a wor...	psa please drink pepsi soda worker company put...	#psa pleas do not drink ani pepsi soda, a work...
3	deep-fried left wings demo-crab cakes barack-a...	deepfried left wings democrab cakes barackamol...	deep-fri left wing demo-crab cake barack-amol ...
4	42 million dead in bloodiest black friday week...	42 million dead bloodiest black friday weekend...	42 million dead in bloodiest black friday week...
5	42 million dead in bloodiest black friday week...	42 million dead bloodiest black friday weekend...	42 million dead in bloodiest black friday week...
6	#prayforchristopher 5k run .. well we are walk...	prayforchristopher 5k run well walking lol sav...	#prayforchristoph 5k run .. well we are walk l...
7	a photo of black nurses saving the life of a k...	photo black nurses saving life kkk member	a photo of black nurs save the life of a kkk m...
8	a photo of black nurses saving the life of a k...	photo black nurses saving life kkk member	a photo of black nurs save the life of a kkk m...
9	a photo of black nurses saving the life of a k...	photo black nurses saving life kkk member	a photo of black nurs save the life of a kkk m...

In [68]:
```python
from nltk.stem.snowball import SnowballStemmer
SnowballStemmer.languages
```

Out[68]:
```
('arabic',
 'danish',
 'dutch',
 'english',
 'finnish',
 'french',
 'german',
 'hungarian',
 'italian',
 'norwegian',
 'porter',
 'portuguese',
 'romanian',
 'russian',
 'spanish',
 'swedish')
```

In [69]:
```python
#lemminizer
lemmatizer = WordNetLemmatizer()
def lemmatize_words(TWEET):
    return " ".join([lemmatizer.lemmatize(word) for word in TWEET.split()])

df["TWEET_lemmatized"] = df["TWEET"].apply(lambda TWEET: lemmatize_words(TWEET))
df.head()
```

Out[69]:

	TWEET	TWEET_stopfreq	TWEET_stemmed	TWEET_lemmatized
0	seriously? racist mcdonaldâ€™s sign is obvious...	seriously racist mcdonaldâ€™s sign obviously hoax	seriously? racist mcdonaldâ€™ sign is obvious ...	seriously? racist mcdonaldâ€™s sign is obvious...
1	hoax: mcdonald's issues official statement on ...	hoax mcdonalds issues official statement racis...	hoax: mcdonald' issu offici statement on racis...	hoax: mcdonald's issue official statement on r...
2	#psa please do not drink any pepsi soda, a wor...	psa please drink pepsi soda worker company put...	#psa pleas do not drink ani pepsi soda, a work...	#psa please do not drink any pepsi soda, a wor...
3	deep-fried left wings demo-crab cakes barack-a...	deepfried left wings democrab cakes barackamol...	deep-fri left wing demo-crab cake barack-amol ...	deep-fried left wing demo-crab cake barack-amo...
4	42 million dead in bloodiest black friday week...	42 million dead bloodiest black friday weekend...	42 million dead in bloodiest black friday week...	42 million dead in bloodiest black friday week...

In [70]:
```python
lemmatizer.lemmatize("sleeping")
```

Out[70]:
```
'sleeping'
```

In [71]:
```python
lemmatizer.lemmatize("sleeping","v") # v1
```

Out[71]:
```
'sleep'
```

In [72]:
```python
print("The Word is : stripes")
print("Lemma result for verb : ",lemmatizer.lemmatize("stripes", 'v'))
print("Lemma result for noun : ",lemmatizer.lemmatize("stripes", 'n'))
```

```
The Word is : stripes
Lemma result for verb :  strip
Lemma result for noun :  stripe
```

In [73]:
```python
def remove_urls(TWEET):
    url_pattern = re.compile(r'https?://\S+|www\.\S+')
    return url_pattern.sub(r'', TWEET)
```

In [74]:
```python
s = 'a\tb\nc\td'
print(s)
```

```
a       b
c       d
```

In [75]:
```python
s = r'a\tb\nc\td'
print(s)
```

```
a\tb\nc\td
```

In [76]:
```python
#removal url
TWEET1 = "This is my website, https://www.abc.com, check it out"
remove_urls(TWEET1)
```

Out[76]:
```
'This is my website,  check it out'
```

In [77]:
```python
TWEET = "Want to learn more. Checkout www.h2o.ai for additional information"
remove_urls(TWEET)
```

Out[77]:
```
'Want to learn more. Checkout  for additional information'
```

In [78]:
```python
import numpy as np
import pandas as pd
import re  # used for preprocessing
import nltk  # Natural Language Toolkit, used for preprocessing
import string #used for preprocessing
from nltk.stem.porter import PorterStemmer
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords

stops = set(stopwords.words('english'))
```

In [79]:
```python
df_whole = pd.read_csv("D:/AI/tweets.csv", nrows=2308)
df = df_whole[["TWEET"]]
df_whole.head()
```

Out[79]:

	TWEET_ID	TWEET	TWEET_Label
0	80080680482123700	seriously? racist mcdonaldâ€™s sign is obvious...	false
1	80084555733803000	hoax: mcdonald's issues official statement on ...	false
2	91728807081426900	#psa please do not drink any pepsi soda, a wor...	false
3	265953285247209000	deep-fried left wings demo-crab cakes barack-a...	false
4	273182568298450000	42 million dead in bloodiest black friday week...	false

In [80]:
```python
#df = df_whole.iloc[:,1:2]
#df["TWEET"] = df["TWEET"].astype(str)
#df.head()
```

In [81]:
```python
# remove all urls
def remove_urls(TWEET):
    url_pattern = re.compile(r'https?://\S+|www\.\S+')
    return url_pattern.sub(r'', TWEET)
# make all text lowercase
def TWEET_lowercase(TWEET):
    return TWEET.lower()
# remove numbers
def remove_numbers(TWEET):
    result = re.sub(r'\d+', '', TWEET)
    return result
# remove punctuation
def remove_punctuation(TWEET):
    translator = str.maketrans('', '', string.punctuation)
    return TWEET.translate(translator)
# tokenize
def tokenize(TWEET):
    TWEET = word_tokenize(TWEET)
    return TWEET
# remove stopwords
stop_words = set(stopwords.words('english'))
def remove_stopwords(TWEET):
    TWEET = [i for i in TWEET if not i in stop_words]
    return TWEET
# lemmatize
lemmatizer = WordNetLemmatizer()
def lemmatize(TWEET):
    TWEET = [lemmatizer.lemmatize(token) for token in TWEET]
    return TWEET

def preprocessing(TWEET):
    TWEET = TWEET_lowercase(TWEET)
    TWEET = remove_urls(TWEET)
    TWEET = remove_numbers(TWEET)
    TWEET = remove_punctuation(TWEET)
    TWEET = tokenize(TWEET)
    TWEET = remove_stopwords(TWEET)
    TWEET = lemmatize(TWEET)
    TWEET = ' '.join(TWEET)
    return TWEET
```

In [82]:
```python
df.head()
```

Out[82]:

	TWEET
0	seriously? racist mcdonaldâ€™s sign is obvious...
1	hoax: mcdonald's issues official statement on ...
2	#psa please do not drink any pepsi soda, a wor...
3	deep-fried left wings demo-crab cakes barack-a...
4	42 million dead in bloodiest black friday week...

In [83]:
```python
pp_TWEET_train = [] # our preprocessed text column
for TWEET_data in df['TWEET']:
    pp_TWEET_data = preprocessing(TWEET_data)
    pp_TWEET_train.append(pp_TWEET_data)
df['pp_TWEET'] = pp_TWEET_train # add the preprocessed text as a column
```

```
<ipython-input-83-70a74f6f9d0c>:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['pp_TWEET'] = pp_TWEET_train # add the preprocessed text as a column
```

In [84]:
```python
df.head()
```

Out[84]:

	TWEET	pp_TWEET
0	seriously? racist mcdonaldâ€™s sign is obvious...	seriously racist mcdonaldâ€™s sign obviously h...
1	hoax: mcdonald's issues official statement on ...	hoax mcdonalds issue official statement racist...
2	#psa please do not drink any pepsi soda, a wor...	psa please drink pepsi soda worker company put...
3	deep-fried left wings demo-crab cakes barack-a...	deepfried left wing democrab cake barackamole ...
4	42 million dead in bloodiest black friday week...	million dead bloodiest black friday weekend re...

In [85]:
```python
final_TWEET_data = list(df['pp_TWEET'])
```

In [86]:
```python
from sklearn.feature_extraction.text import TfidfVectorizer

tf=TfidfVectorizer()

# the vectorizer must be fit onto the entire corpus
fitted_vectorizer = tf.fit(final_TWEET_data)

transform_all = fitted_vectorizer.transform(df['pp_TWEET'])
```

In [87]:
```python
print(transform_all)
```

```
  (0, 5079)     0.06521763769677442
  (0, 4358)     0.3840850279674288
  (0, 4271)     0.39177991493899084
  (0, 3849)     0.3840850279674288
  (0, 3361)     0.46397169887280154
  (0, 3032)     0.44060634547925404
  (0, 2283)     0.3712261460816987
  (1, 5079)     0.054772357129968016
  (1, 5070)     0.35611582655700963
  (1, 4566)     0.3290323013102362
  (1, 4358)     0.32256983023391905
  (1, 4272)     0.38966182288010026
  (1, 3849)     0.32256983023391905
  (1, 3375)     0.24231715876571722
  (1, 3031)     0.35611582655700963
  (1, 2529)     0.34531642786961864
  (1, 2283)     0.31177043154652806
  (2, 5296)     0.24261053638144783
  (2, 4453)     0.2873154504280058
  (2, 4102)     0.19188831175394777
  (2, 3824)     0.23364236746402559
  (2, 3792)     0.2873154504280058
  (2, 3623)     0.4529628457632189
  (2, 3550)     0.24811141169880563
  (2, 2453)     0.22337641720438547
  :     :
  (2305, 2344)  0.36781422126680346
  (2305, 1137)  0.36781422126680346
  (2305, 1084)  0.36781422126680346
  (2305, 108)   0.27891894658947497
  (2306, 5372)  0.3287727690933629
  (2306, 5228)  0.2691627288007266
  (2306, 5079)  0.048664263597869135
  (2306, 4875)  0.3287727690933629
  (2306, 4761)  0.23935770865440845
  (2306, 3402)  0.19271927156691673
  (2306, 3269)  0.1844776578790455
  (2306, 3131)  0.3287727690933629
  (2306, 2985)  0.28153292982821004
  (2306, 2677)  0.24309915912390298
  (2306, 2412)  0.29896774894704475
  (2306, 947)   0.20611088771644837
  (2306, 870)   0.3068074947120958
  (2306, 163)   0.3287727690933629
  (2307, 5079)  0.07062351325424504
  (2307, 5008)  0.47712810795962923
  (2307, 2985)  0.40857177590373817
  (2307, 2677)  0.3527951604970869
  (2307, 2412)  0.43387387826984286
  (2307, 947)   0.29911631111426457
  (2307, 870)   0.4452512288760943
```