Python code for removing of short words like is,a etc

In [ ]:
```
import pandas as pd
import numpy as np

data= pd.read_csv("dataset.csv")
data.head
```

Out[ ]:
```
<bound method NDFrame.head of                                                    Text  Label
0      Damn I thought they had strict gun laws in Ger...      0
1      I dont care about what it stands for or anythi...      0
2                        It's not a group it's an idea lol      0
3                               So it's not just America!      0
4      The dog is a spectacular dancer considering he...      0
...                                                  ...    ...
17591    I find rats nicer and cleaner than most Chinese      1
17592  Check out this niggar, they hit things like wi...      1
17593  this country has become an absolute shambles, ...      0
17594  Me aged 16 = anti-Semitism is bad Me aged 18 =...      1
17595    so messed up saying blacks don't deserve rights      0

[17596 rows x 2 columns]>
```

In [ ]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17596 entries, 0 to 17595
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   Text    17596 non-null  object
 1   Label   17596 non-null  int64 
dtypes: int64(1), object(1)
memory usage: 275.1+ KB
```

After observing the dataset, we can infer that we need to clean and transform the raw text data into a format suitable for our model. This involves several sub-steps:

1. Lowercasing: Convert all text to lowercase to ensure uniformity, as the model should treat "Hate" and "hate" as the same word.

In [ ]:
```
def text_lower(text):
    return text.lower()
data['Text'] = data['Text'].apply(text_lower)
data.head(20)
```

Out[ ]:

	Text	Label
0	damn i thought they had strict gun laws in ger...	0
1	i dont care about what it stands for or anythi...	0
2	its not a group its an idea lol	0
3	so its not just america	0
4	the dog is a spectacular dancer considering he...	0
5	if ppl dont wear masks you complain if ppl wea...	0
6	we should send them all back to africa	1
7	checking to see if its whataboutism from the p...	0
8	as a european i approve this	0
9	idk which of these groups to join theyre all s...	0
10	sometimes the messenger is just as important a...	0
11	this is the future we all want	0
12	wow the monkey calls someone else a monkey lol	1
13	they are the gypsies of america	1
14	i know that guy	0
15	meal team six	0
16	will the guy respawn	0
17	im sorry did they misspell texas s	0
18	their combined gold in wow is amazing though	0
19	what a chad wearing a cowboy hat to his wedding	0

2. Removing Punctuation: Strip out punctuation to focus on the words themselves.

In [ ]:
```
import re
def remove_punctuation(text):
    return re.sub(r'[^\w\s]', '', text)

data['Text'] = data['Text'].apply(remove_punctuation)
data.head
```

Out[ ]:
```
<bound method NDFrame.head of                                                 Text  Label
0      damn i thought they had strict gun laws in ger...      0
1      i dont care about what it stands for or anythi...      0
2                        its not a group its an idea lol      0
3                                so its not just america      0
4      the dog is a spectacular dancer considering he...      0
...                                                  ...    ...
17591    i find rats nicer and cleaner than most chinese      1
17592  check out this niggar they hit things like wil...      1
17593  this country has become an absolute shambles t...      0
17594  me aged 16  antisemitism is bad me aged 18  an...      1
17595     so messed up saying blacks dont deserve rights      0

[17596 rows x 2 columns]>
```

3. Remove URLs: URLs do not give any information when we try to analyze text from words.

In [ ]:
```
def remove_url(text):
    return re.sub(r'http\S+|www\S+|https\S+', '', text, flags=re.MULTILINE)

data['Text'] = data['Text'].apply(remove_url)
```

In [ ]:
```
data.head
```

Out[ ]:
```
<bound method NDFrame.head of                                                 Text  Label
0      damn i thought they had strict gun laws in ger...      0
1      i dont care about what it stands for or anythi...      0
2                        its not a group its an idea lol      0
3                                so its not just america      0
4      the dog is a spectacular dancer considering he...      0
...                                                  ...    ...
17591    i find rats nicer and cleaner than most chinese      1
17592  check out this niggar they hit things like wil...      1
17593  this country has become an absolute shambles t...      0
17594  me aged 16  antisemitism is bad me aged 18  an...      1
17595     so messed up saying blacks dont deserve rights      0

[17596 rows x 2 columns]>
```

4. Removing short words like is, a

In [ ]:
```
import pandas as pd
data  = pd.read_csv("dataset.csv")
data['Text'] = data['Text'].apply(lambda x: ' '.join([word for word in x.split() if len(word) > 3]))
data
```

Out[ ]:

	Text	Label
0	Damn thought they strict laws Germany	0
1	dont care about what stands anything connected...	0
2	It's group it's idea	0
3	it's just America!	0
4	spectacular dancer considering left feet.	0
...	...	...
17591	find rats nicer cleaner than most Chinese	1
17592	Check this niggar, they things like wild people	1
17593	this country become absolute shambles, amount ...	0
17594	aged anti-Semitism aged anti-Semitism doesn't ...	1
17595	messed saying blacks don't deserve rights	0

17596 rows × 2 columns

5. Stemming

In [ ]:
```
import pandas as pd
data  = pd.read_csv("dataset.csv")
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize
stemmer = PorterStemmer()

data['Text'] = [' '.join([stemmer.stem(word) for word in word_tokenize(text)]) for text in data['Text']]
data
```

Out[ ]:

	Text	Label
0	damn i thought they had strict gun law in germani	0
1	i dont care about what it stand for or anyth i...	0
2	it 's not a group it 's an idea lol	0
3	so it 's not just america !	0
4	the dog is a spectacular dancer consid he ha t...	0
...	...	...
17591	i find rat nicer and cleaner than most chines	1
17592	check out thi niggar , they hit thing like wil...	1
17593	thi countri ha becom an absolut shambl , the a...	0
17594	me age 16 = anti-semit is bad me age 18 = anti...	1
17595	so mess up say black do n't deserv right	0

17596 rows × 2 columns