

3107- JAWAHAR ENGINEERING COLLEGE

S.No.	NaanMudhalvan ID	Position	Name	Department
		Faculty Mentor	V.Nivaskumar	
1.	au310721106001	Team Head	Junia Susheela Shalom	ECE
2.	au310721205001	Team member	M.Hariragavan	IT
3.	au310721205002	Team Member	S.Muthunivas Pandi	IT

Subject Title- AI 101- Artificial Intelligence

Project Title-Building a Smarter AI-Powered Spam Classifier: Phase-3

Junia03 /
NMIBM

Code

Issues

Pull requests

Actions

Projects

Wiki

Security

Insights

Set

NMIBM / Project.ipynb



Junia03 Add files via upload

14 hours ago



2425 lines (2425 loc) · 311 KB

Preview

Code

Blame

Raw



```
In [1]: import numpy as np
import pandas as pd
```

```
In [2]: dataset = pd.read_csv('spam.csv')
```

```
In [3]: dataset.sample(5)
```

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
1246	ham I do know what u mean, is the king of not hav...		NaN	NaN	NaN
2339	ham Cheers for the message Zogtorius. Iâ€ve been s...		NaN	NaN	NaN
3340	ham Still i have not checked it da...		NaN	NaN	NaN
1351	ham Yo theres no class tmrw right?		NaN	NaN	NaN
2997	ham No b4 Thursday		NaN	NaN	NaN

```
In [4]: dataset.shape
```

```
Out[4]: (5572, 5)
```

```
#1.Data Cleaning
#2.EDA
#3.Text Preprocessing
#4.Model Building
#5.Evaluation
#6.Improvement
#7.Website
#8.Deploy
```

1.Data Cleaning

```
In [6]: dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype  
 --- 
 0   v1          5572 non-null   object 
 1   v2          5572 non-null   object 
 2   Unnamed: 2   50 non-null    object 
 3   Unnamed: 3   12 non-null    object 
 4   Unnamed: 4   6 non-null     object 
dtypes: object(5)
memory usage: 217.8+ KB
```

```
In [7]: #drop last three columns
dataset.drop(columns=['Unnamed: 2','Unnamed: 3','Unnamed: 4'],inplace=True)
```

```
In [8]: dataset.sample(5)
```

	v1	v2
--	----	----

```
1934 ham          Hope you are having a great day.
196 ham          Did u got that persons story
3346 ham          No problem baby. Is this is a good time to tal...
2049 ham          How much is blackberry bold2 in nigeria.
1220 spam         No. 1 Nokia Tone 4 ur mob every week! Just txt...
```

In [9]:

```
#rename the columns
dataset.rename(columns={'v1':'type', 'v2':'text'}, inplace=True)
dataset.sample(5)
```

Out[9]:

	type	text
1273	spam	network operator. The service is free. For T &...
3946	ham	Sorry, went to bed early, nightnight
1440	ham	Cool breeze... Bright sun... Fresh flower... T...
3519	ham	No it will reach by 9 only. She telling she wi...
1722	ham	Thought praps you meant another one. Goodo! I'...

In [10]:

```
from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()
```

In [11]:

```
encoder.fit_transform(dataset['type'])
```

Out[11]:

```
array([0, 0, 1, ..., 0, 0, 0])
```

In [12]:

```
dataset['type'] = encoder.fit_transform(dataset['type'])
dataset.head()
```

Out[12]:

	type	text
0	0	Go until jurong point, crazy.. Available only ...
1	0	Ok lar... Joking wif u oni...
2	1	Free entry in 2 a wkly comp to win FA Cup fina...
3	0	U dun say so early hor... U c already then say...
4	0	Nah I don't think he goes to usf, he lives aro...

In [13]:

```
#missing values
dataset.isnull().sum()
```

Out[13]:

```
type      0
text      0
dtype: int64
```

In [14]:

```
#check for duplicate values
dataset.duplicated().sum()
```

Out[14]:

```
403
```

In [15]:

```
#remove duplicates
dataset.drop_duplicates(keep='first')
```

Out[15]:

	type	text
0	0	Go until jurong point, crazy.. Available only ...
1	0	Ok lar... Joking wif u oni...
2	1	Free entry in 2 a wkly comp to win FA Cup fina...
3	0	U dun say so early hor... U c already then say...
4	0	Nah I don't think he goes to usf, he lives aro...
...
5567	1	This is the 2nd time we have tried 2 contact u...
5568	0	Will l_b going to esplanade fr home?
5569	0	Pity, * was in mood for that. So...any other s...
5570	0	The guy did some bitching but I acted like i'd...
5571	0	Rofl. Its true to its name

5169 rows × 2 columns

In [16]:

```
dataset = dataset.drop_duplicates(keep='first')
```

In [17]:

```
dataset.duplicated().sum()
```

Out[17]: 0

In [18]:

```
dataset.shape
```

Out[18]: (5169, 2)

2.EDA

In [19]:

```
dataset.head()
```

Out[19]:

	type	text
0	0	Go until jurong point, crazy.. Available only ...
1	0	Ok lar... Joking wif u oni...
2	1	Free entry in 2 a wkly comp to win FA Cup fina...
3	0	U dun say so early hor... U c already then say...
4	0	Nah I don't think he goes to usf, he lives aro...

In [20]:

```
dataset.value_counts()
```

Out[20]: type

0 &lt;#&gt; in mca. But not conform.

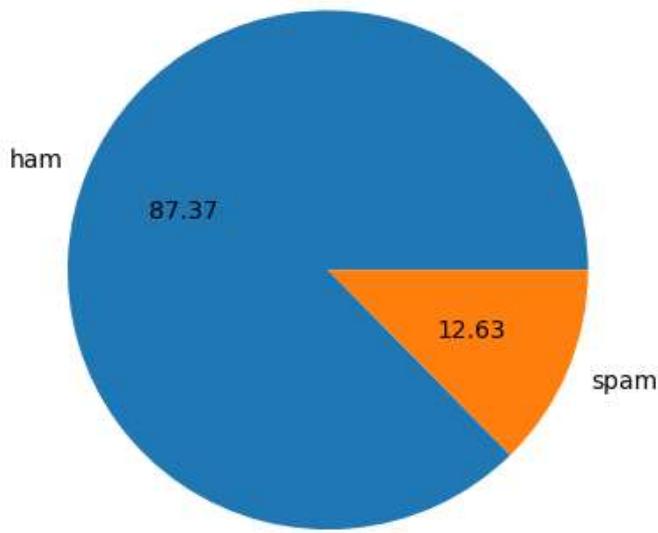
```
1 Thats cool. i liked your photos. You are very sexy!
1 That's good, because I need drugs
1 That's fine, have him give me a call if he knows what he wants or has any questions
1 That's fine, I'll bitch at you about it later then
1

..
I want to send something that can sell fast. &lt;#&gt; k is not easy money.
1 I want to see your pretty pussy...
1 I want to lick your pussy now...
1 I want to go to perumbavoor
1 we tried to contact you re your response to our offer of a new nokia fone and camcorder
hit reply or call 08000930705 for delivery 1
Name: count, Length: 5169, dtype: int64
```

```
In [21]: dataset['type'].value_counts()
```

```
Out[21]: type
0    4516
1    653
Name: count, dtype: int64
```

```
In [22]: import matplotlib.pyplot as plt
plt.pie(dataset['type'].value_counts(), labels=['ham', 'spam'], autopct="%0.2f")
plt.show()
```



```
In [23]: #Data is imbalanced
```

```
In [24]: import nltk
```

```
In [25]: nltk.download('punkt')
```

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

Out[25]: True

In [26]: `dataset['text'].apply(len)`

```
Out[26]: 0      111
1      29
2      155
3      49
4      61
...
5567    161
5568    37
5569    57
5570    125
5571    26
Name: text, Length: 5169, dtype: int64
```

In [27]: `dataset['num_characters'] = dataset['text'].apply(len)`
`dataset.head()`

	type	text	num_characters
0	0	Go until jurong point, crazy.. Available only ...	111
1	0	Ok lar... Joking wif u oni...	29
2	1	Free entry in 2 a wkly comp to win FA Cup fina...	155
3	0	U dun say so early hor... U c already then say...	49
4	0	Nah I don't think he goes to usf, he lives aro...	61

In [28]: `# num of words`
`dataset['text'].apply(lambda x:len(nltk.word_tokenize(x)))`

```
Out[28]: 0      24
1      8
2      37
3      13
4      15
...
5567    35
5568    9
5569    15
5570    27
5571    7
Name: text, Length: 5169, dtype: int64
```

In [29]: `dataset['num_words']= dataset['text'].apply(lambda x:len(nltk.word_tokenize(x)))`

In [30]: `dataset.head()`

	type	text	num_characters	num_words
0	0	Go until jurong point, crazy.. Available only ...	111	24
1	0	Ok lar... Joking wif u oni...	29	8

2	1	Free entry in 2 a wkly comp to win FA Cup fina...	155	37
3	0	U dun say so early hor... U c already then say...	49	13
4	0	Nah I don't think he goes to usf, he lives aro...	61	15

In [31]: `dataset['text'].apply(lambda x:len(nltk.sent_tokenize(x)))`

Out[31]:

0	2
1	2
2	2
3	1
4	1
	..
5567	4
5568	1
5569	2
5570	1
5571	2
	Name: text, Length: 5169, dtype: int64

In [32]: `dataset['num_sentences'] = dataset['text'].apply(lambda x:len(nltk.sent_tokenize(x)))`
`dataset.head()`

Out[32]:

	type	text	num_characters	num_words	num_sentences
0	0	Go until jurong point, crazy.. Available only ...	111	24	2
1	0	Ok lar... Joking wif u oni...	29	8	2
2	1	Free entry in 2 a wkly comp to win FA Cup fina...	155	37	2
3	0	U dun say so early hor... U c already then say...	49	13	1
4	0	Nah I don't think he goes to usf, he lives aro...	61	15	1

In [33]: `dataset[['num_characters', 'num_words', 'num_sentences']].describe()`

Out[33]:

	num_characters	num_words	num_sentences
count	5169.000000	5169.000000	5169.000000
mean	78.977945	18.455794	1.965564
std	58.236293	13.324758	1.448541
min	2.000000	1.000000	1.000000
25%	36.000000	9.000000	1.000000
50%	60.000000	15.000000	1.000000
75%	117.000000	26.000000	2.000000
max	910.000000	220.000000	38.000000

In [34]: `#ham`
`dataset[dataset['type'] == 0][['num_characters', 'num_words', 'num_sentences']].describe()`

Out[34]:

	num_characters	num_words	num_sentences
count	4516.000000	4516.000000	4516.000000

mean	70.459256	17.123782	1.820195
std	56.358207	13.493970	1.383657
min	2.000000	1.000000	1.000000
25%	34.000000	8.000000	1.000000
50%	52.000000	13.000000	1.000000
75%	90.000000	22.000000	2.000000
max	910.000000	220.000000	38.000000

In [35]:

```
#spam  
dataset[dataset['type'] == 1][['num_characters', 'num_words', 'num_sentences']].describe()
```

Out[35]:

	num_characters	num_words	num_sentences
count	653.000000	653.000000	653.000000
mean	137.891271	27.667688	2.970904
std	30.137753	7.008418	1.488425
min	13.000000	2.000000	1.000000
25%	132.000000	25.000000	2.000000
50%	149.000000	29.000000	3.000000
75%	157.000000	32.000000	4.000000
max	224.000000	46.000000	9.000000

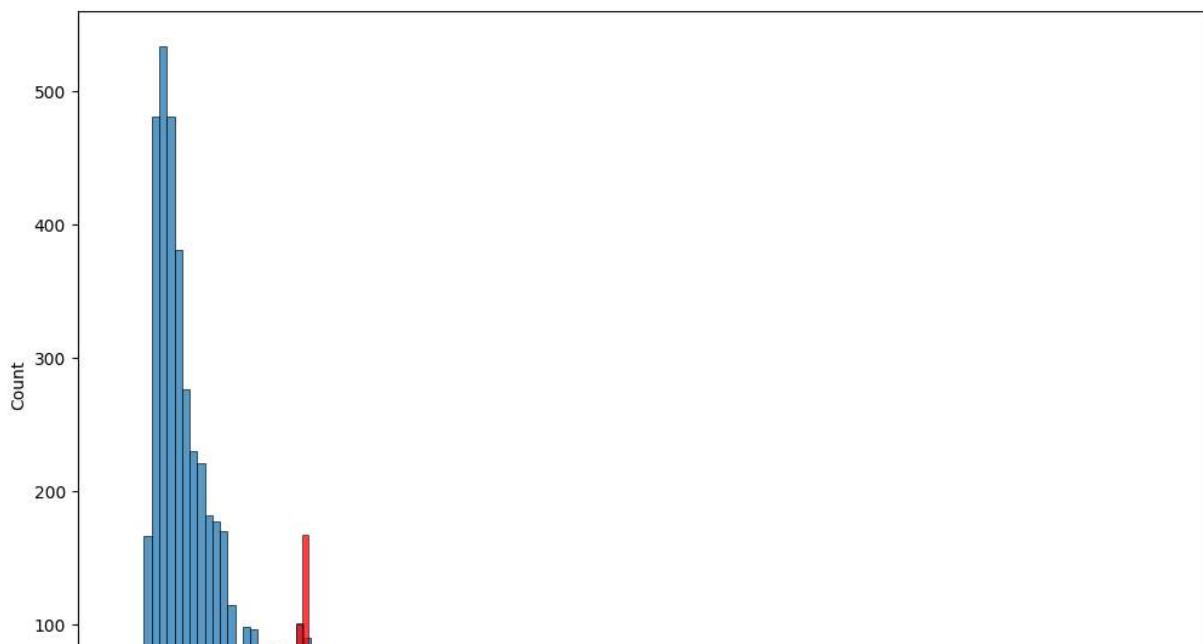
In [36]:

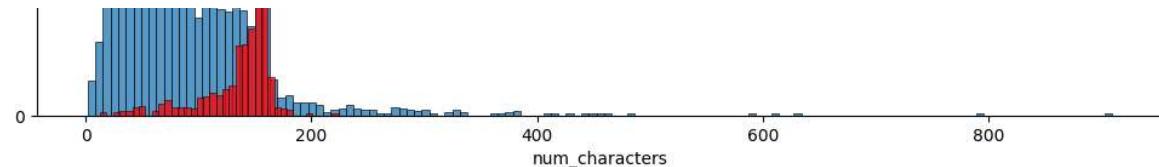
```
import seaborn as sns
```

In [37]:

```
plt.figure(figsize = (12,8))  
sns.histplot(dataset[dataset['type'] == 0]['num_characters'])  
sns.histplot(dataset[dataset['type'] == 1]['num_characters'], color='red')
```

Out[37]: <Axes: xlabel='num_characters', ylabel='Count'>

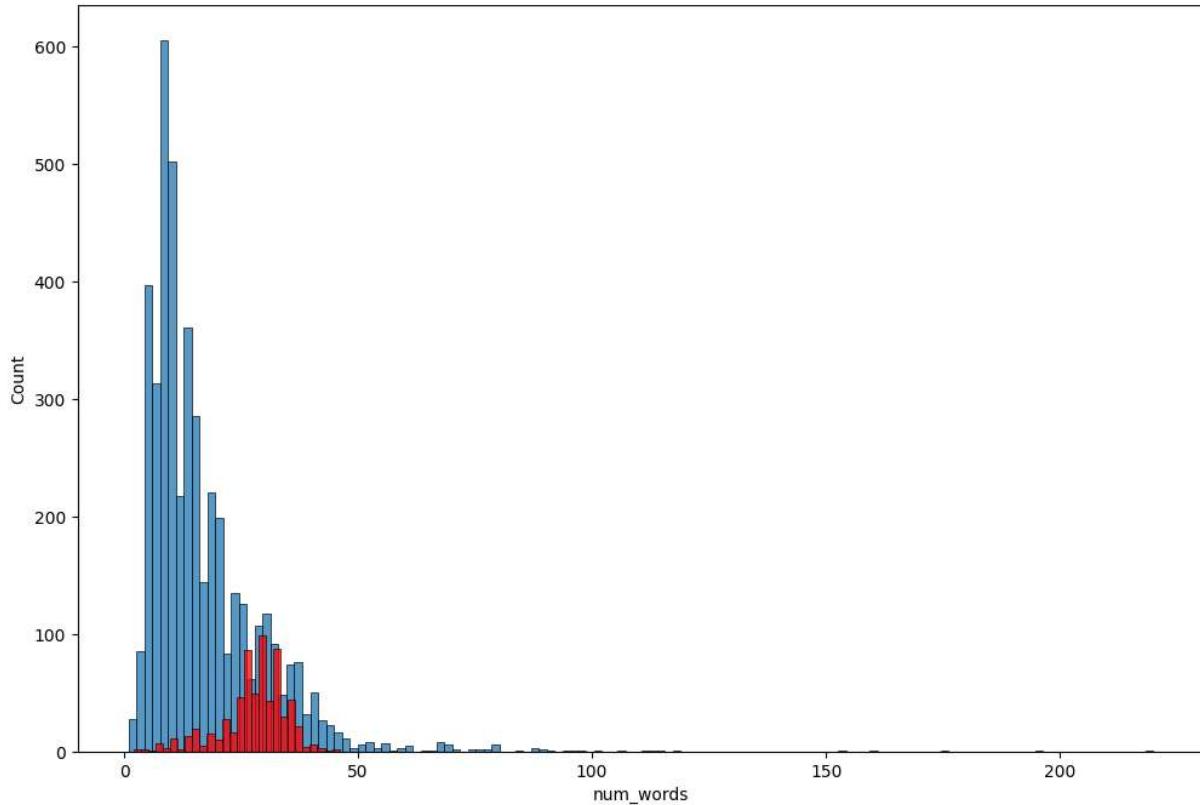




In [38]:

```
plt.figure(figsize = (12,8))
sns.histplot(dataset[dataset['type'] == 0]['num_words'])
sns.histplot(dataset[dataset['type'] == 1]['num_words'],color='red')
```

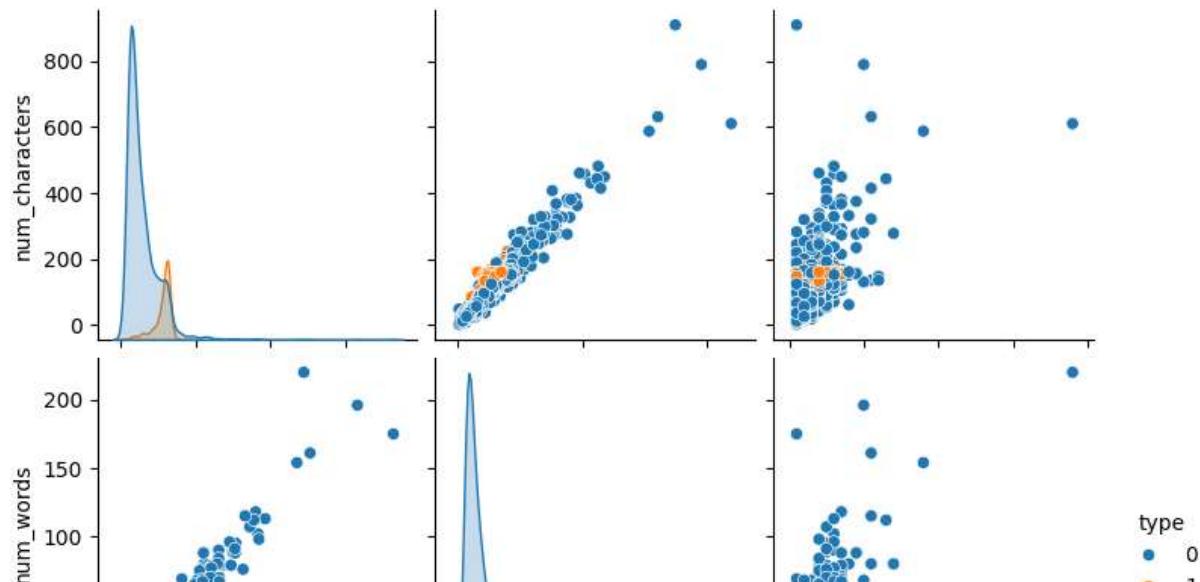
Out[38]: <Axes: xlabel='num_words', ylabel='Count'>

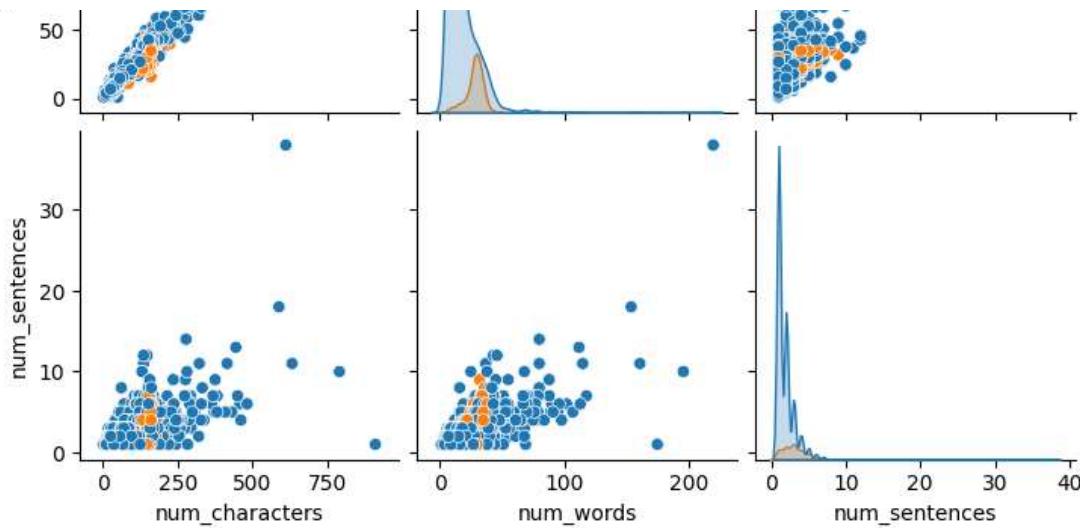


In [39]:

```
sns.pairplot(dataset, hue = 'type')
```

Out[39]: <seaborn.axisgrid.PairGrid at 0x241040d6ed0>





3. Data Preprocessing

```
In [40]: def transform_text(text):
    text = text.lower()
    text = nltk.word_tokenize(text)

    y=[]
    for i in text:
        if i.isalnum():
            y.append(i)

    return y
```

```
In [41]: transform_text('Hi How Are You %%')
```

```
Out[41]: ['hi', 'how', 'are', 'you']
```

```
In [42]: dataset['text'][2000]
```

```
Out[42]: "But i'll b going 2 sch on mon. My sis need 2 take smth."
```

```
In [43]: from nltk.corpus import stopwords
stopwords.words('english')
```

```
Out[43]: ['i',
'me',
'my',
'myself',
'we',
'our',
'ours',
'ourselves',
'you',
"you're",
"you've",
"you'll",
"you'd",
'your',
'yours',
'yourself',
'yourselves',
'he',
...']
```

```
'him',
'his',
'himself',
'she',
"she's",
'her',
'hers',
'herself',
'it',
"it's",
'its',
'itself',
'they',
'them',
'their',
'theirs',
'themselves',
'what',
'which',
'who',
'whom',
'this',
'that',
"that'll",
'these',
'those',
'am',
'is',
'are',
'was',
'were',
'be',
'veen',
'being',
'have',
'has',
'had',
'having',
'do',
'does',
'did',
'doing',
'a',
'an',
'the',
'and',
'but',
'if',
'or',
'because',
'as',
'until',
'while',
'of',
'at',
'by',
'for',
'with',
'about',
'against',
'between',
'into',
'through',
'during',
'before',
'after',
'above',
'below',
'to',
```

```
'from',
'up',
'down',
'in',
'out',
'on',
'off',
'over',
'under',
'again',
'further',
'then',
'once',
'here',
'there',
'when',
'where',
'why',
'how',
'all',
'any',
'both',
'each',
'few',
'more',
'most',
'other',
'some',
'such',
'no',
'nor',
'not',
'only',
'own',
'same',
'so',
'than',
'too',
'very',
's',
't',
'can',
'will',
'just',
'don',
"don't",
'should',
"should've",
'now',
'd',
'll',
'm',
'o',
're',
've',
'y',
'ain',
'aren',
"aren't",
'couldn',
"couldn't",
'didn',
"didn't",
'doesn',
"doesn't",
'hadn',
"hadn't",
'hasn',
"hasn't",
```

```
'haven',
'haven\'t',
'isn',
"isn't",
'ma',
'mightn',
"mightn't",
'mustn',
"mustn't",
'needn',
"needn't",
'shan',
"shan't",
'shouldn',
"shouldn't",
'wasn',
"wasn't",
'weren',
"weren't",
>won',
>won't",
>wouldn',
>wouldn't']
```

In [44]: `import string`

In [45]: `string.punctuation`

Out[45]: `'!"#$%&\'()*+,.-./:;<=>?@[\\]^_`{|}~'`

In [54]: `from nltk.stem.porter import PorterStemmer
ps = PorterStemmer()
ps.stem('dancing')`

Out[54]: `'danc'`

In [56]: `def transform_text(text):
 text = text.lower()
 text = nltk.word_tokenize(text)

 y=[]
 for i in text:
 if i.isalnum():
 y.append(i)
 text = y[:]
 y.clear()

 for i in text:
 if i not in stopwords.words('english') and i not in string.punctuation:
 y.append(i)

 text = y[:]
 y.clear()

 for i in text:
 y.append(ps.stem(i))

 return " ".join(y)`

In [57]: `transform_text('I love the lectures on machine learning')`

Out[57]: `'love lectur machin learn'`

In [58]: `dataset['text'][10]`

Out[58]: "I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today."

In [59]: `dataset['transformed_text'] = dataset['text'].apply(transform_text)`

In [60]: `dataset.head()`

	type	text	num_characters	num_words	num_sentences	transformed_text
0	0	Go until jurong point, crazy.. Available only ...	111	24	2	go jurong point crazi avail bugi n great world...
1	0	Ok lar... Joking wif u oni...	29	8	2	ok lar joke wif u oni
2	1	Free entry in 2 a wkly comp to win FA Cup fina...	155	37	2	free entri 2 wkli comp win fa cup final tkt 21...
3	0	U dun say so early hor... U c already then say...	49	13	1	u dun say earli hor u c alreadi say
4	0	Nah I don't think he goes to usf, he lives aro...	61	15	1	nah think goe usf live around though

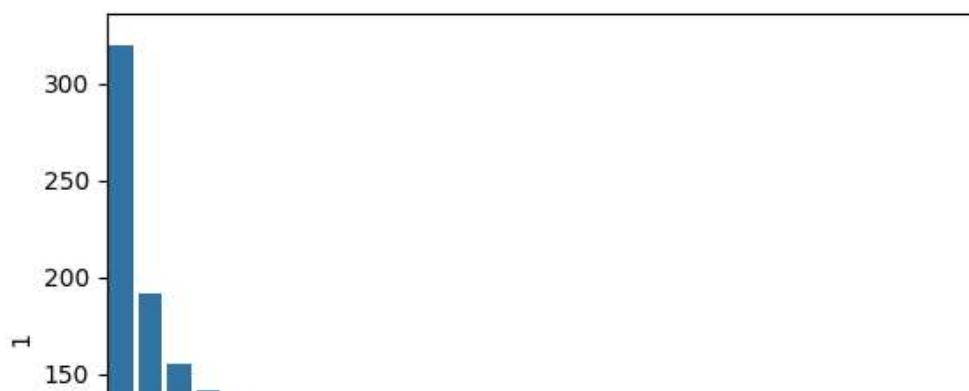
In [61]: `spam_corpus = []
for msg in dataset[dataset['type'] == 1]['transformed_text'].tolist():
 for word in msg.split():
 spam_corpus.append(word)`

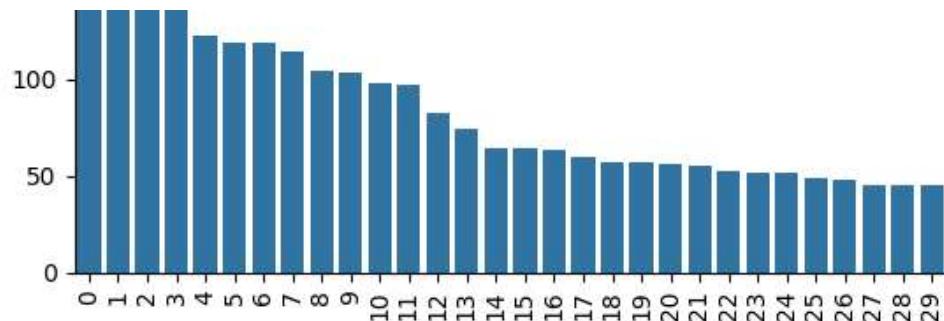
In [64]: `len(spam_corpus)`

Out[64]: 9939

In [82]: `from collections import Counter
sns.barplot(pd.DataFrame(Counter(spam_corpus).most_common(30))[1])
plt.xticks(rotation='vertical')
plt.show`

Out[82]: <function matplotlib.pyplot.show(close=None, block=None)>





In [84]:

```
ham_corpus = []
for msg in dataset[dataset['type'] == 1]['transformed_text'].tolist():
    for word in msg.split():
        ham_corpus.append(word)
```

In [85]:

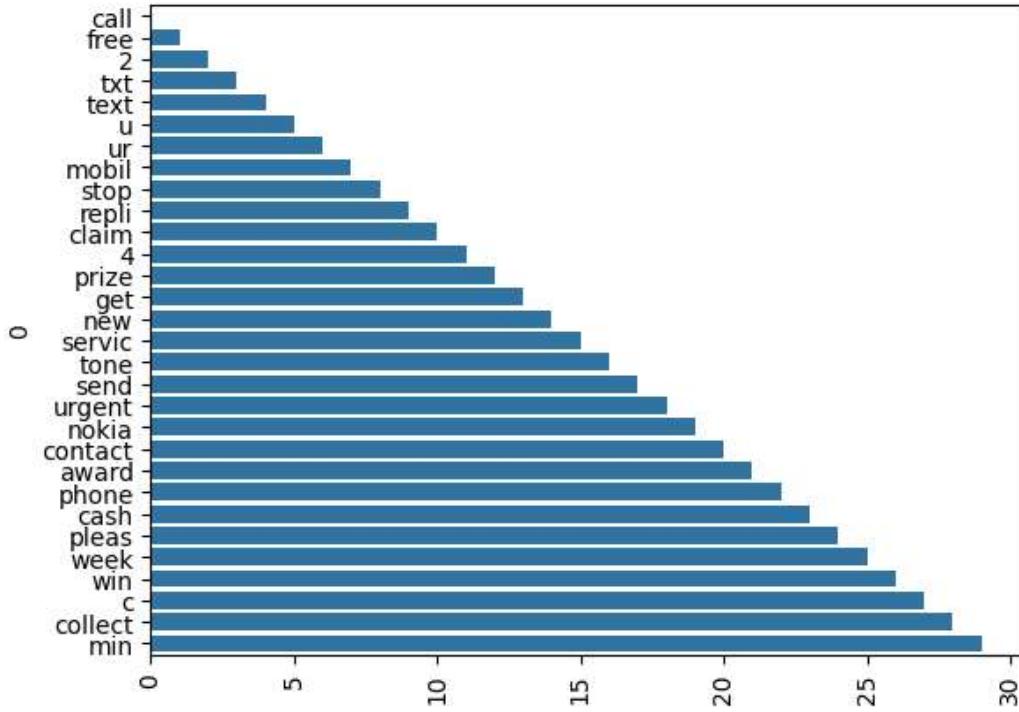
```
len(ham_corpus)
```

Out[85]: 9939

In [87]:

```
from collections import Counter
sns.barplot(pd.DataFrame(Counter(ham_corpus).most_common(30))[0])
plt.xticks(rotation='vertical')
plt.show
```

Out[87]: <function matplotlib.pyplot.show(close=None, block=None)>



In []:

