In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import nltk
import seaborn as sns
import string
```

In [2]:
```
df=pd.read_csv('spam.csv',encoding='latin-1')
df.head()
```

Out[2]:

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN

In [3]:
```
df=df.drop(columns=['Unnamed: 2','Unnamed: 3','Unnamed: 4'])
```

In [4]:
```
df.v1.value_counts()
```

Out[4]:
```
ham     4825
spam     747
Name: v1, dtype: int64
```

In [5]:
```
df.shape
```

Out[5]:
```
(5572, 2)
```

In [6]:
```
for i in range(0,5572):
    if df['v1'][i]=="spam":
        df['v1'][i]=1
    else:
        df['v1'][i]=0
df['v1']
```

Out[6]:
```
0       0
1       0
2       1
3       0
4       0
       ..
5567    1
5568    0
5569    0
5570    0
5571    0
Name: v1, Length: 5572, dtype: object
```

In [7]:
```
df.v1.value_counts()
```

Out[7]:
```
0    4825
1     747
Name: v1, dtype: int64
```

In [8]:
```
df.rename(columns={'v1':'target','v2':'text'},inplace=True)
df.sample(10)
```

Out[8]:

	target	text
4854	0	Same to u...
3035	0	;-) ok. I feel like john lennon.
1582	0	Yep, at derek's house now, see you Sunday <3
820	0	Good afternoon starshine! How's my boytoy? Doe...
2117	0	Wish u many many returns of the day.. Happy bi...
4321	0	Sorry . I will be able to get to you. See you ...
5507	0	I want to be inside you every night...
3611	0	K, my roommate also wants a dubsack and anothe...
1978	0	No I'm in the same boat. Still here at my moms...
535	0	Good afternoon, my love! How goes that day ? I...

In [9]:
```
df.duplicated().sum()
```

Out[9]:
```
403
```

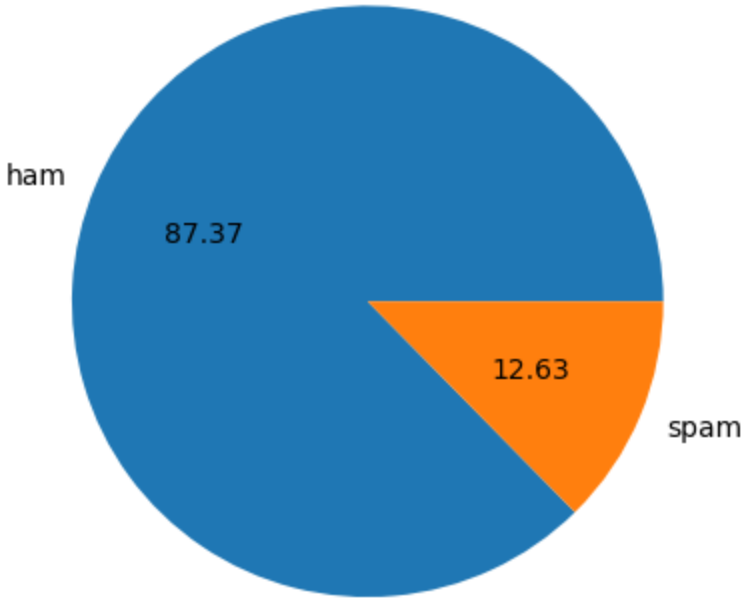
In [10]:
```
df=df.drop_duplicates(keep='first')
df.duplicated().sum()
```

Out[10]:
```
0
```

In [11]:
```
df.shape
```

Out[11]:
```
(5169, 2)
```

In [12]:
```
plt.pie(df['target'].value_counts(),labels=['ham','spam'],autopct="%0.2f")
plt.show()
```



In [13]:
```
nltk.download('punkt')
```

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

Out[13]:
```
True
```

In [14]:
```
df['num_char']=df['text'].apply(len)
```

In [15]:
```
df['num_word']=df['text'].apply(lambda x:len(nltk.word_tokenize(x)))
```

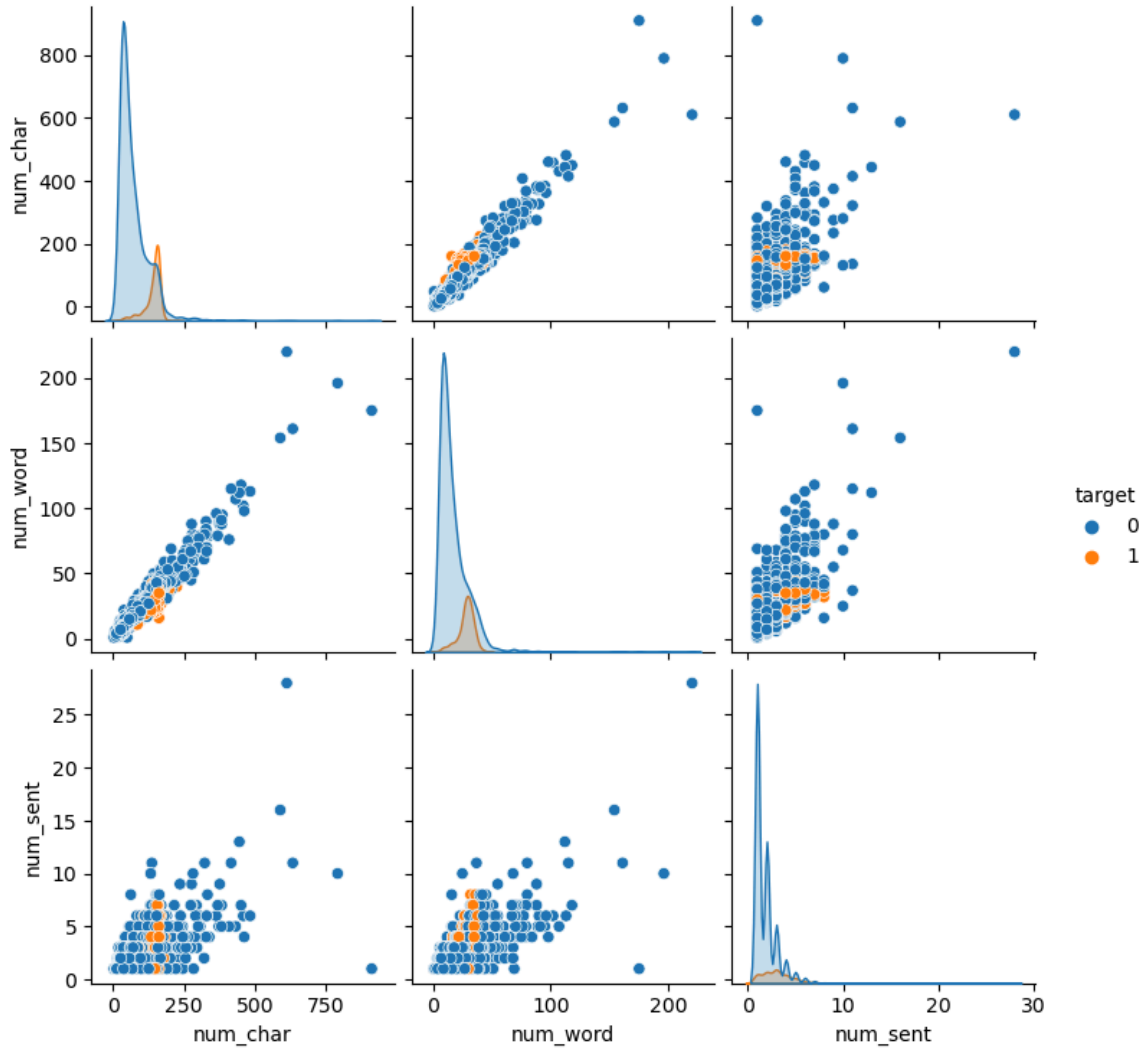
In [16]:
```
df['num_sent']=df['text'].apply(lambda x:len(nltk.sent_tokenize(x)))
```

In [18]:
```
plt.figure(figsize=(50,10))
sns.histplot(df[df['target']==0]['num_char'])
sns.histplot(df[df['target']==1]['num_char'], color='green')

plt.figure(figsize=(50,10))
sns.histplot(df[df['target']==0]['num_char'])
sns.histplot(df[df['target']==1]['num_char'], color='orange')

plt.figure(figsize=(50,10))
sns.histplot(df[df['target']==0]['num_sent'])
sns.histplot(df[df['target']==1]['num_sent'], color='pink')
```

Out[18]: <Axes: xlabel='num_sent', ylabel='Count'>

In [19]:
```
plt.figure(figsize=(30,10))
sns.pairplot(df,hue='target')
```

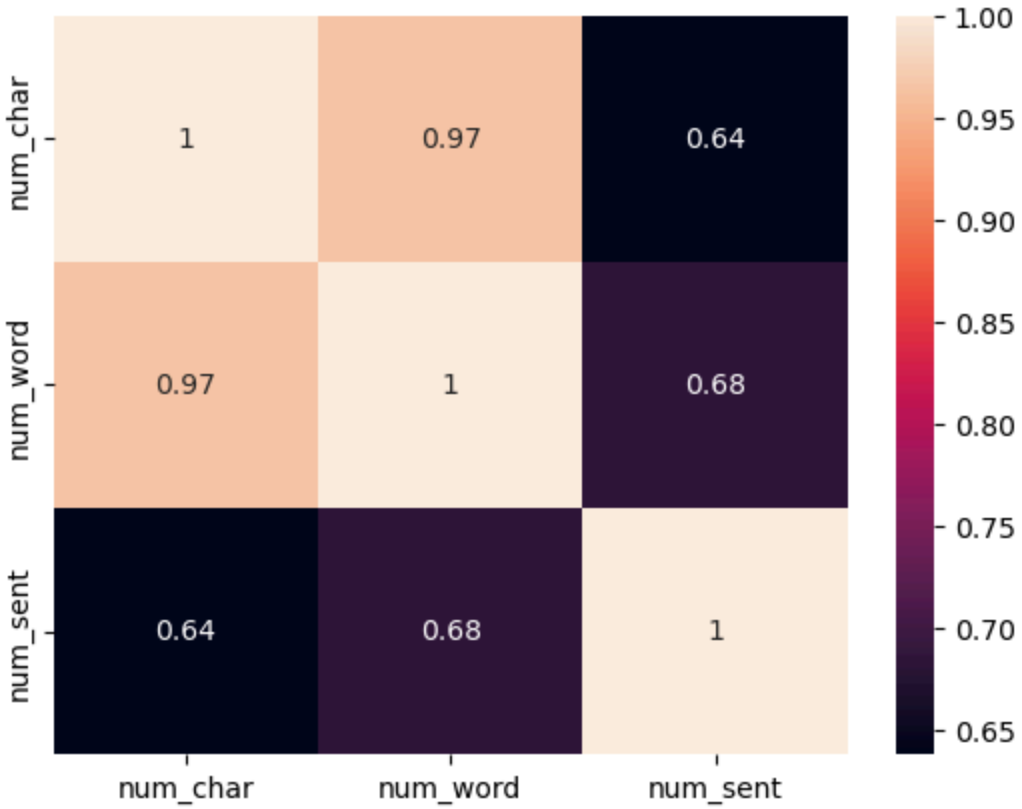
Out[19]:
```
<seaborn.axisgrid.PairGrid at 0x22111848bd0>

<Figure size 3000x1000 with 0 Axes>
```



In [20]:
```
sns.heatmap(df.corr(),annot = True)
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_18124\2221401063.py:1: FutureW
arning: The default value of numeric_only in DataFrame.corr is deprecated.
In a future version, it will default to False. Select only valid columns or
specify the value of numeric_only to silence this warning.
  sns.heatmap(df.corr(),annot = True)
```

Out[20]: <Axes: >



In [21]:
```
df.head()
```

Out[21]:

	target	text	num_char	num_word	num_sent
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 [22]:
```
df.describe()
```

Out[22]:

	num_char	num_word	num_sent
count	5169.000000	5169.000000	5169.000000
mean	78.977945	18.453279	1.947185
std	58.236293	13.324793	1.362406
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	28.000000

In [23]:
```
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
ps = PorterStemmer()
ps.stem('Sleeping')
```

Out[23]:
```
'sleep'
```

In [24]:
```
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.punctuati
            y.append(i)
    text=y[:]
    y.clear()

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

    return " ".join(y)
```

In [25]:
```
df.head(10)
```

Out[25]:

	target	text	num_char	num_word	num_sent
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
5	1	FreeMsg Hey there darling it's been 3 week's n...	148	39	4
6	0	Even my brother is not like to speak with me. ...	77	18	2
7	0	As per your request 'Melle Melle (Oru Minnamin...	160	31	2
8	1	WINNER!! As a valued network customer you have...	158	32	5
9	1	Had your mobile 11 months or more? U R entitle...	154	31	3

In [ ]:
```
```