In [14]:
```
import numpy as np
import pandas as pd
```

In [15]:
```
df=pd.read_csv('spam.csv')
print(df)
```

	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
5	spam	FreeMsg Hey there darling it's been 3 week's n...
6	ham	Even my brother is not like to speak with me. ...
7	ham	As per your request 'Melle Melle (Oru Minnamin...
8	spam	WINNER!! As a valued network customer you have...
9	spam	Had your mobile 11 months or more? U R entitle...
10	ham	I'm gonna be home soon and i don't want to tal...
11	spam	SIX chances to win CASH! From 100 to 20,000 po...
12	spam	URGENT! You have won a 1 week FREE membership ...
13	ham	I've been searching for the right words to tha...
14	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
15	spam	XXXMobileMovieClub: To use your credit, click ...
16	ham	Oh k...i'm watching here:)
17	ham	Eh u remember how 2 spell his name... Yes i di...
18	ham	Fine if that□s the way u feel. That□s the way ...
19	spam	England v Macedonia - dont miss the goals/team...
20	ham	Is that seriously how you spell his name?
21	ham	I‘m going to try for 2 months ha ha only joking
22	ham	So ü pay first lar... Then when is da stock co...
23	ham	Aft i finish my lunch then i go str down lor. ...
24	ham	Ffffffffff. Alright no way I can meet up with ...
25	ham	Just forced myself to eat a slice. I'm really ...
26	ham	Lol your always so convincing.
27	ham	Did you catch the bus ? Are you frying an egg ...
28	ham	I'm back & we're packing the car now, I'll...
29	ham	Ahhh. Work. I vaguely remember that! What does...
...	...	...
5542	ham	Armand says get your ass over to epsilon
5543	ham	U still havent got urself a jacket ah?
5544	ham	I'm taking derek & taylor to walmart, if I...
5545	ham	Hi its in durban are you still on this number
5546	ham	Ic. There are a lotta childporn cars then.
5547	spam	Had your contract mobile 11 Mnths? Latest Moto...
5548	ham	No, I was trying it all weekend ;V
5549	ham	You know, wot people wear. T shirts, jumpers, ...
5550	ham	Cool, what time you think you can get here?
5551	ham	Wen did you get so spiritual and deep. That's ...
5552	ham	Have a safe trip to Nigeria. Wish you happines...
5553	ham	Hahaha..use your brain dear
5554	ham	Well keep in mind I've only got enough gas for...
5555	ham	Yeh. Indians was nice. Tho it did kane me off ...
5556	ham	Yes i have. So that's why u texted. Pshew...mi...
5557	ham	No. I meant the calculation is the same. That ...
5558	ham	Sorry, I'll call later
5559	ham	if you aren't here in the next <#> hou...
5560	ham	Anything lor. Juz both of us lor.
5561	ham	Get me out of this dump heap. My mom decided t...
5562	ham	Ok lor... Sony ericsson salesman... I ask shuh...
5563	ham	Ard 6 like dat lor.
5564	ham	Why don't you wait 'til at least wednesday to ...
5565	ham	Huh y lei...
5566	spam	REMINDER FROM O2: To get 2.50 pounds free call...
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will ü b going to esplanade fr home?
5569	ham	Pity, * was in mood for that. So...any other s...
5570	ham	The guy did some bitching but I acted like i'd...

```
5571      ham                          Rofl. Its true to its name

[5572 rows x 2 columns]
```

In [16]:
```
df.sample(5)
```

Out[16]:

	Category	Message
1262	ham	Thank you so much. When we skyped wit kz and s...
4324	ham	Aight well keep me informed
2414	ham	Lol please do. Actually send a pic of yourself...
4613	ham	Sorry da. I gone mad so many pending works wha...
989	ham	Yun ah.the ubi one say if ü wan call by tomorr...

In [17]:
```
df.shape
```

Out[17]:
```
(5572, 2)
```

In [18]:
```
#DATA CLEANING
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 2 columns):
Category    5572 non-null object
Message     5572 non-null object
dtypes: object(2)
memory usage: 87.1+ KB
```

In [19]:
```
#RENAMING THE COLUMNS
df.rename(columns={'Category':'target','Message':'text'},inplace=True)
df.sample(5)
```

Out[19]:

	target	text
3384	ham	K... Must book a not huh? so going for yoga ba...
2792	ham	… we r stayin here an extra week, back next we...
2067	ham	Then. You are eldest know.
3913	spam	You have an important customer service announc...
878	spam	Sunshine Quiz Wkly Q! Win a top Sony DVD playe...

In [20]:
```
from sklearn.preprocessing import LabelEncoder
encoder=LabelEncoder()
```

In [21]:
```
df['target']=encoder.fit_transform(df['target'])
```

In [22]:
```
df.head()
```

Out[22]:

	target	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 [23]:
```
#MISSING VALUES
df.isnull().sum()
```

Out[23]:
```
target    0
text      0
dtype: int64
```

In [24]:
```
#CHECK FOR DUPLICATED VALUES
df.duplicated().sum()
```

Out[24]:
```
415
```

In [25]:
```
#REMOVE DUPLICATES
df=df.drop_duplicates(keep='first')
```

In [26]:
```
df.duplicated().sum()
```

Out[26]:
```
0
```

In [27]:
```
df.shape
```

Out[27]:
```
(5157, 2)
```

In [28]:
```
#EXPLORATIVE DATA ANALYSIS(EDA)
df.head()
```

Out[28]:

	target	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 [29]:
```
df['target'].value_counts()
```

Out[29]:
```
0    4516
1     641
Name: target, dtype: int64
```

In [30]:
```
import matplotlib.pyplot as plt
plt.pie(df['target'].value_counts(),labels=['ham','spam'],autopct="%0.2f")
plt.show()
```

```
<Figure size 640x480 with 1 Axes>
```

In [31]:
```
#DATA IS IMBALANCED
import nltk
```

In [32]:
```
!pip install nltk
```

```
Requirement already satisfied: nltk in c:\users\sk nazeer pasha\anaconda3
\lib\site-packages (3.3)
Requirement already satisfied: six in c:\users\sk nazeer pasha\anaconda3\l
ib\site-packages (from nltk) (1.11.0)

distributed 1.21.8 requires msgpack, which is not installed.
You are using pip version 10.0.1, however version 21.3.1 is available.
You should consider upgrading via the 'python -m pip install --upgrade pi
p' command.
```

In [33]:
```
nltk.download('punkt')
```

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

Out[33]:
```
True
```

In [34]:
```
df['num_characters']=df['text'].apply(len)
```

In [35]:
```
df.head()
```

Out[35]:

	target	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 [36]:
```
df['num_words']=df['text'].apply(lambda x:len(nltk.word_tokenize(x)))
```

In [37]:
```
df.head()
```

Out[37]:

	target	text	num_characters	num_words
0	0	Go until jurong point, crazy.. Available only ...	111	23
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 [38]:
```
df['num_sentences']=df['text'].apply(lambda x:len(nltk.sent_tokenize(x)))
```

In [39]:
```
df.head()
```

Out[39]:

	target	text	num_characters	num_words	num_sentences
0	0	Go until jurong point, crazy.. Available only ...	111	23	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 [40]:
```
df[['num_characters','num_words','num_sentences']].describe()
```

Out[40]:

	num_characters	num_words	num_sentences
count	5157.000000	5157.000000	5157.000000
mean	79.103936	18.390537	1.965290
std	58.382922	13.307527	1.439549
min	2.000000	1.000000	1.000000
25%	36.000000	9.000000	1.000000
50%	61.000000	15.000000	1.000000
75%	118.000000	26.000000	2.000000
max	910.000000	219.000000	38.000000

In [41]:
```
#ham
df[df['target']==0][['num_characters','num_words','num_sentences']].describe
```

Out[41]:

	num_characters	num_words	num_sentences
count	4516.000000	4516.000000	4516.000000
mean	70.869353	17.101417	1.822852
std	56.708301	13.488402	1.374848
min	2.000000	1.000000	1.000000
25%	34.000000	8.000000	1.000000
50%	53.000000	13.000000	1.000000
75%	91.000000	22.000000	2.000000
max	910.000000	219.000000	38.000000

In [42]:
```
#spam
df[df['target']==1][['num_characters','num_words','num_sentences']].describe
```

Out[42]:

	num_characters	num_words	num_sentences
count	641.000000	641.000000	641.000000
mean	137.118565	27.472699	2.968799
std	30.399707	6.988134	1.486069
min	7.000000	2.000000	1.000000
25%	130.000000	25.000000	2.000000
50%	148.000000	29.000000	3.000000
75%	157.000000	32.000000	4.000000
max	223.000000	44.000000	9.000000

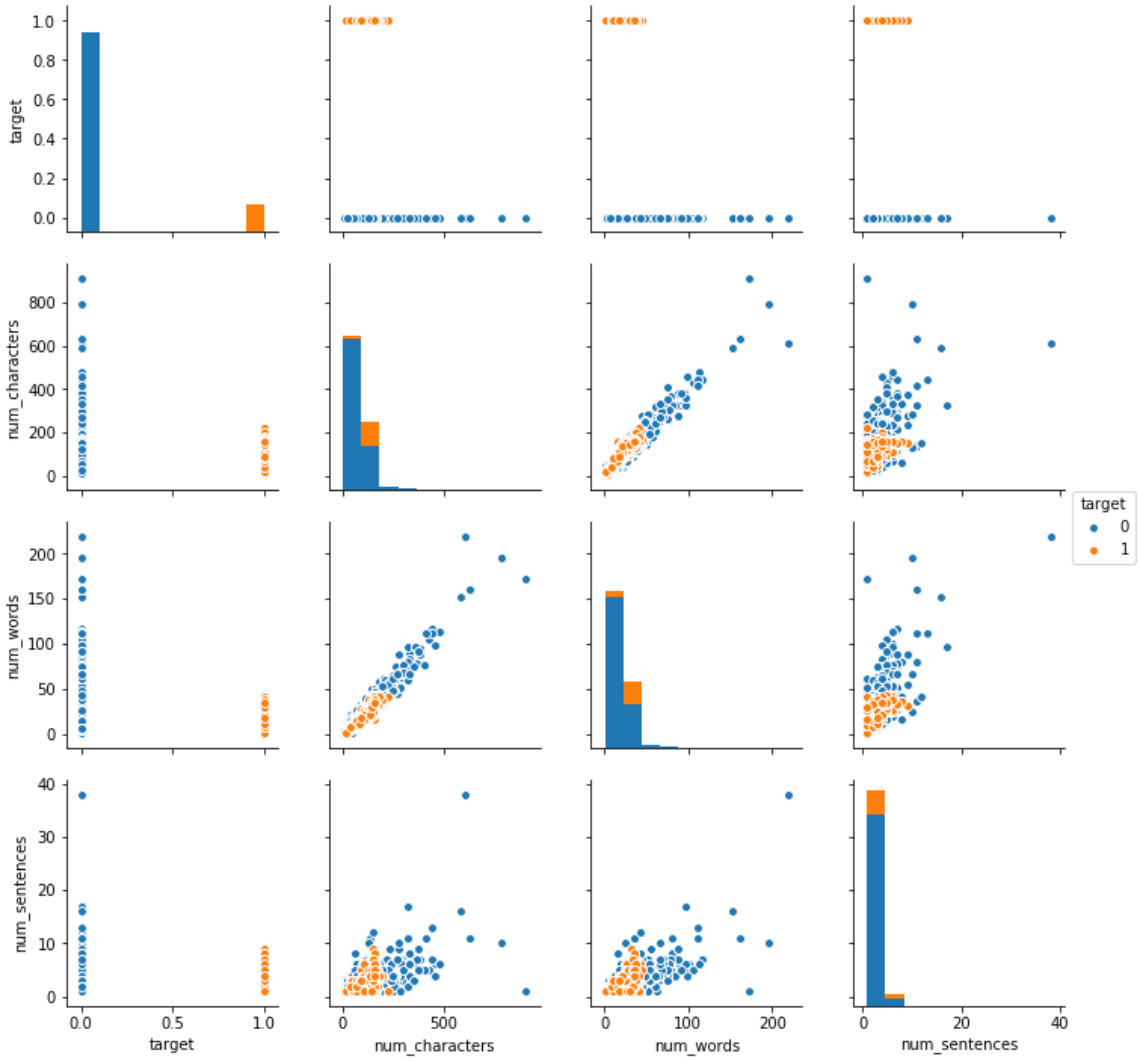
In [43]:
```
import seaborn as sns
```

In [ ]:
```

```

In [44]:
```
sns.pairplot(df,hue='target')
```

Out[44]:
```
<seaborn.axisgrid.PairGrid at 0x14326799b00>
```



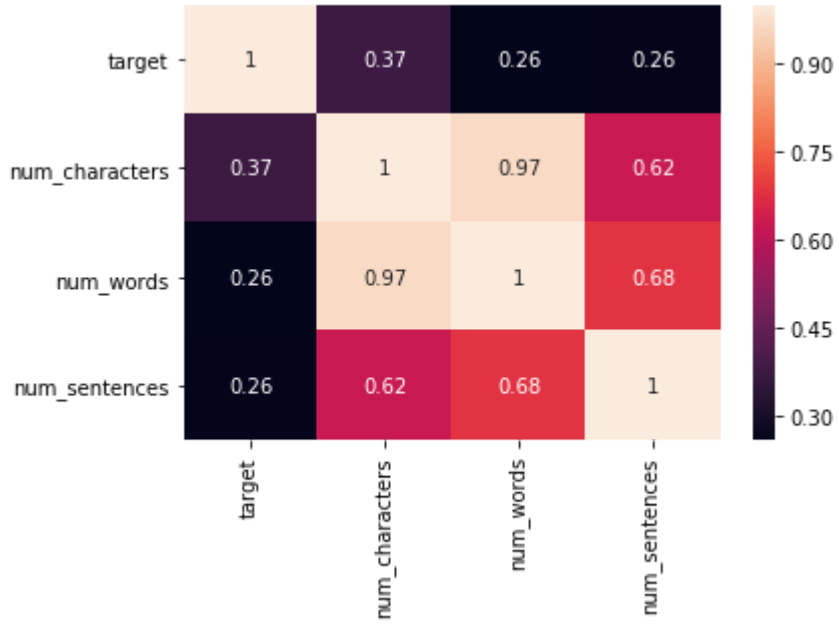
In [45]:
```
df.corr()
```

Out[45]:

	target	num_characters	num_words	num_sentences
target	1.000000	0.374409	0.257150	0.262657
num_characters	0.374409	1.000000	0.965669	0.624267
num_words	0.257150	0.965669	1.000000	0.682739
num_sentences	0.262657	0.624267	0.682739	1.000000

In [46]:
```
sns.heatmap(df.corr(),annot=True)
```

Out[46]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x143275a3cf8>
```



Data Preprocessing

Lower Case

Tokenization

Removing special characters

Removing stopwords and punctuation

Stemming

In [47]:
```
def transform_text(text):
    text=text.lower()
    return text
```

In [ ]:
```

```

In [ ]:
```

```

In [48]:
```
transform_text('I Am Khajabi')
```

Out[48]:
```
'i am khajabi'
```

In [49]:
```
def transform_text(text):
    text=text.lower()
    text=nltk.word_tokenize(text)
    return text
```

In [50]:
```
transform_text('I Am Khajabi')
```

Out[50]:
```
['i', 'am', 'khajabi']
```

In [51]:
```
df['text'][1000]
```

Out[51]:
```
'Aight will do, thanks again for comin out'
```

In [52]:
```
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 [53]:
```
transform_text('I Am Khajabi 20% eg')
```

Out[53]:
```
['i', 'am', 'khajabi', '20', 'eg']
```

In [54]:
```
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 [55]:
```
import nltk
nltk.download('stopwords')
```

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

Out[55]:
```
True
```

In [56]:
```
from nltk.corpus import stopwords
stopwords.words('english')
```

Out[56]:
```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
```

In [57]:
```
import string
string.punctuation
```

Out[57]:
```
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

In [58]:
```
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)
    return y
```

In [59]:
```
transform_text('I Am khajabi')
```

Out[59]:
```
['khajabi']
```

In [60]:
```
from nltk.stem.porter import PorterStemmer
ps=PorterStemmer()
ps.stem('loving')
```

Out[60]:
```
'love'
```

In [61]:
```
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 [62]:
```
transform_text('happining')
```

Out[62]:
```
'happin'
```

In [63]:
```
transform_text('Aight will do, thanks again for comin out')
```

Out[63]:
```
'aight thank comin'
```

In [64]:
```
df['transform_text']= df['text'].apply(transform_text)
```

In [65]:
```
df.head()
```

Out[65]:

	target	text	num_characters	num_words	num_sentences	transform_text
0	0	Go until jurong point, crazy.. Available only ...	111	23	2	go jurong point avail bugi n great world la e ...
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 [ ]:
```

```

In [ ]:
```

```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: