In [22]:
```
import pandas as pd
import numpy as np
```

BEFORE PREPROCESSING

In [23]:
```
f = pd.read_csv('D:\Project\smsspamcollection\SMSSpamCollection.txt', delimite
r = "\t")
print(f)
```

```
       ham  \
0      ham
1     spam
2      ham
3      ham
4     spam
...    ...
5566  spam
5567   ham
5568   ham
5569   ham
5570   ham

     Go until jurong point, crazy.. Available only in bugis n great world la
e buffet... Cine there got amore wat...
0                          Ok lar... Joking wif u oni...
1     Free entry in 2 a wkly comp to win FA Cup fina...
2     U dun say so early hor... U c already then say...
3     Nah I don't think he goes to usf, he lives aro...
4     FreeMsg Hey there darling it's been 3 week's n...
...                                                 ...
5566  This is the 2nd time we have tried 2 contact u...
5567              Will ü b going to esplanade fr home?
5568  Pity, * was in mood for that. So...any other s...
5569  The guy did some bitching but I acted like i'd...
5570                       Rofl. Its true to its name

[5571 rows x 2 columns]
```

In [24]:
```
'''df=pd.read_table('D:\Project\smsspamcollection\SMSSpamCollection.txt',heade
r=None)
print(df)
f.to_csv('Spamtest.csv')'''
```

Out[24]:
```
"df=pd.read_table('D:\\Project\\smsspamcollection\\SMSSpamCollection.txt',hea
der=None)\nprint(df)\nf.to_csv('Spamtest.csv')"
```

In [25]:
```
df=pd.read_csv('Spamtest.csv')
df
```

Out[25]:

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

5572 rows × 2 columns

In [26]:
```
df=df.replace(['ham','spam'],[0,1])
df
```

Out[26]:

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

5572 rows × 2 columns

In [27]:
```
# Total ham(0) and spam(1) messages
df['type'].value_counts()
```

Out[27]:
```
0    4825
1     747
Name: type, dtype: int64
```

In [28]:
```
df['Count']=0
for i in np.arange(0,len(df.text)):
    df.loc[i,'Count'] = len(df.loc[i,'text'])
df
```

Out[28]:

	type	text	Count
0	0	Ok lar... Joking wif u oni...	29
1	1	Free entry in 2 a wkly comp to win FA Cup fina...	155
2	0	U dun say so early hor... U c already then say...	49
3	0	Nah I don't think he goes to usf, he lives aro...	61
4	1	FreeMsg Hey there darling it's been 3 week's n...	147
...	...	...	...
5567	0	Will ü b going to esplanade fr home?	36
5568	0	Pity, * was in mood for that. So...any other s...	57
5569	0	The guy did some bitching but I acted like i'd...	125
5570	0	Rofl. Its true to its name	26
5571	0	Go until jurong point, crazy.. Available only ...	111

5572 rows × 3 columns

In [29]:
```
df1=pd.read_csv('sms_spam.csv')
df1
```

Out[29]:

	type	text
0	ham	Hope you are having a good week. Just checking in
1	ham	K..give back my thanks.
2	ham	Am also doing in cbe only. But have to pay.
3	spam	complimentary 4 STAR Ibiza Holiday or £10,000 ...
4	spam	okmail: Dear Dave this is your final notice to...
...	...	...
5554	ham	You are a great role model. You are giving so ...
5555	ham	Awesome, I remember the last time we got someb...
5556	spam	If you don't, your prize will go to another cu...
5557	spam	SMS. ac JSco: Energy is high, but u may not kn...
5558	ham	Shall call now dear having food

5559 rows × 2 columns

In [31]:
```
df1=df1.replace(['ham','spam'],[0,1])
df1
```

Out[31]:

	type	text
0	0	Hope you are having a good week. Just checking in
1	0	K..give back my thanks.
2	0	Am also doing in cbe only. But have to pay.
3	1	complimentary 4 STAR Ibiza Holiday or £10,000 ...
4	1	okmail: Dear Dave this is your final notice to...
...	...	...
5554	0	You are a great role model. You are giving so ...
5555	0	Awesome, I remember the last time we got someb...
5556	1	If you don't, your prize will go to another cu...
5557	1	SMS. ac JSco: Energy is high, but u may not kn...
5558	0	Shall call now dear having food

5559 rows × 2 columns

In [34]:
```
df1['Count']=0
for i in np.arange(0,len(df1.text)):
    df1.loc[i,'Count'] = len(df1.loc[i,'text'])
df1
```

Out[34]:

	type	text	Count
0	0	Hope you are having a good week. Just checking in	49
1	0	K..give back my thanks.	23
2	0	Am also doing in cbe only. But have to pay.	43
3	1	complimentary 4 STAR Ibiza Holiday or £10,000 ...	149
4	1	okmail: Dear Dave this is your final notice to...	161
...	...	...	...
5554	0	You are a great role model. You are giving so ...	245
5555	0	Awesome, I remember the last time we got someb...	88
5556	1	If you don't, your prize will go to another cu...	145
5557	1	SMS. ac JSco: Energy is high, but u may not kn...	154
5558	0	Shall call now dear having food	31

5559 rows × 3 columns

In [37]:
```
df2=pd.concat([df,df1]).drop_duplicates().reset_index(drop=True)
```

In [38]:
```
# Total ham(0) and spam(1) messages
df2['type'].value_counts()
```

Out[38]:
```
0    5142
1     672
Name: type, dtype: int64
```

AFTER PREPROCESSING

In [39]:
```
df2
```

Out[39]:

	type	text	Count
0	0	Ok lar... Joking wif u oni...	29
1	1	Free entry in 2 a wkly comp to win FA Cup fina...	155
2	0	U dun say so early hor... U c already then say...	49
3	0	Nah I don't think he goes to usf, he lives aro...	61
4	1	FreeMsg Hey there darling it's been 3 week's n...	147
...	...	...	...
5809	0	Dude. What's up. How Teresa. Hope you have bee...	324
5810	1	I want some! My hubby's away, I need a real ma...	157
5811	0	Same here, but I consider walls and bunkers an...	154
5812	0	Ok lor thanx... u in school?	28
5813	0	And stop wondering wow is she ever going to st...	133

5814 rows × 3 columns

In [ ]:
```
```