Spam filter using Naive Bayes algorithm

by Susan Fisher

The purpose of this project is to build a spam filter that classifies text or SMS messages as spam or not spam. The filter will use multinomial Naive Bayes algorithm, which is based on conditional probability. The desired accuracy is greater than 80%.

To train the algorithm, data containing previously assembled SMS messages is used. The data contains 5,572 SMS messages that have been classified by humans. It was assembled by Tiago A. Almeida and Jose Maria Gomez Hidalgo, and can be downloaded from The UCI Machine Learning Repository at:
https://archive.ics.uci.edu/ml/datasets/sms+spam+collection
The following address provides details on the data:
http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/#composition

In [1]:

```
# Read in the csv file as a dataframe, "data.""

import pandas as pd

'''Data does not have a header row.
Parameter, header=None, else 1st row will be assigned as header row'''
data = pd.read_csv('C:/Users/Name/Documents/Python Scripts/DataSets/SMSSpamCollection
',
                   sep='\t',
                   header=None,
                  names=['Label', 'SMS'])
```

Data Exploration

In [2]:

```
data.head()
```

Out[2]:

	Label	SMS
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...

SMS messages are categorized or labeled as "spam" or "ham" for nonspam.

In [3]:

```
data.shape
```

Out[3]:

```
(5572, 2)
```

It's helpful to know the percent of spam and nonspam or "ham" messages in the dataset.

In [4]:

```
# % of spam messages, and % of nonspam messages or "ham"
data['Label'].value_counts(normalize=True)
```

Out[4]:

```
ham     0.865937
spam    0.134063
Name: Label, dtype: float64
```

Split Data

The data will be split into two sets, a training set and a test set.

Eighty percent of the data will be the training set, and it will be used to train the computer how to classify messages.

The remaining twenty percent of the data will be the test set, and used to test how accurately the spam filter classifies new messages.

In [5]:

```
# Randomize dataset to split the data into training set and test set.

'''Parameter, frac=1, randomizes the entire dataset '''
'''Parameter, random_state=1, random numbers that are randomly generated are reproduc
ible'''
randomized = data.sample(frac=1, random_state=1)
```

In [6]:

```
# Training data: computer number of rows
training_index = round( len(randomized) * 0.8 )

# Split randomized data: 80% as Training set and 20% as Testing set
# And reset indices for both datasets
training_data = randomized[:training_index].reset_index(drop=True)
testing = randomized[training_index:].reset_index(drop=True)

print(training_data.shape, testing.shape)
print(testing.head(3))
```

```
(4458, 2) (1114, 2)
  Label                                                SMS
0   ham          Later i guess. I needa do mcat study too.
1   ham             But i haf enuff space got like 4 mb...
2  spam  Had your mobile 10 mths? Update to latest Oran...
```

The percent of spam and "ham" or nonspam messages in both the training and testing datasets should be same as in the original dataset.

In [7]:

```
# Percent of spam and ham messages in training dataset

training_data['Label'].value_counts(normalize=True)
```

Out[7]:

```
ham     0.86541
spam    0.13459
Name: Label, dtype: float64
```

In [8]:

```
# Percent of spam and ham messages in testing dataset

testing['Label'].value_counts(normalize=True)
```

Out[8]:

```
ham     0.868043
spam    0.131957
Name: Label, dtype: float64
```

Data Cleaning - Training dataset

In the Training data set, the "SMS" column, contains the SMS message. In order to manipulate or compare the individual words word in this column, then each row in the column needs to be converted to a list. So only the words, independent of punctuation and case, are considered, then the punctuation needs to be removed and all the words need to be converted to lowercase.

Then each row of the "SMS" column will be converted to a list.

A list of vocabulary words will be created based on the cleaned "SMS" column.

In [9]:

```
# Before cleaning Training data set
training_data.head()
```

Out[9]:

	Label	SMS
0	ham	Yep, by the pretty sculpture
1	ham	Yes, princess. Are you going to make me moan?
2	ham	Welp apparently he retired
3	ham	Havent.
4	ham	I forgot 2 ask ü all smth.. There's a card on ...

In [10]:

```
# "SMS" column: remove punctuation, and make all words lower case

'''To avoid a SettingWithCopy Warning'''
training_data = training_data.copy()

training_data['SMS'] = training_data['SMS'].str.replace('\W', ' ')
training_data['SMS'] = training_data['SMS'].str.lower()

# View some rows to make sure the cleaning worked
training_data.head()
```

Out[10]:

	Label	SMS
0	ham	yep by the pretty sculpture
1	ham	yes princess are you going to make me moan
2	ham	welp apparently he retired
3	ham	havent
4	ham	i forgot 2 ask ü all smth there s a card on ...

In [11]:

```
# SMS column: for every row, split the string into a list of strings

training_data['SMS'] = training_data['SMS'].str.split()
training_data.head(3)
```

Out[11]:

	Label	SMS
0	ham	[yep, by, the, pretty, sculpture]
1	ham	[yes, princess, are, you, going, to, make, me,...
2	ham	[welp, apparently, he, retired]

Create Vocabulary

The vocabulary will be a list of of unique words found in the "SMS" message column of the training dataset.

In [12]:

```
# Training dataset, "SMS" column: create a vocabulary list of unique words

vocabulary = []

for row in training_data['SMS']:
    for word in row:
        vocabulary.append(word)

'''Convert vocabulary to a set to remove any duplicates'''
vocabulary = set(vocabulary)
'''Convert vocabulary back to a list'''
vocabulary = list(vocabulary)

# Number of unique words in vocabulary list and the first ten words.
print(len(vocabulary), vocabulary[:10], sep='\n')
```

```
7783
['silent', 'luv', 'offcampus', 'omg', 'joking', 'treat', 'charlie', '86688', 'vijay',
'jez']
```

Final training dataset

The final training dataset will have a column for each unique word in the vocabulary list. And each row of each word will have a count of each word in the "SMS" message.

To accomplish this, first a dictionary of the unique vocabulary words will be created. The keys will be the words in the vocabulary list. The values will be a list of the count of each key word in every row. Initially the values will be initialized to zero.

Lastly, the training dataset will be combined with this newly created dictionary of unique vocabulary words and word counts.

In [13]:

```
# Create a dictionary of the vocabulary words and training dataset "SMS" column

# Initialize values to 0: for each row, the number of values is the total number of r
ows in the training data.
word_counts_per_sms = {unique_word: [0] * len(training_data['SMS'])
                       for unique_word in vocabulary}

# Populate dictionary values
'''Enumerate method to iterate SMS messages, and each unique_word in vocabulary, or i
```

```python
ndex.
   Outer for loop, iterates through each SMS message.
   Inner for loop, iterates through each word (index) of that SMS message '''
for index, sms in enumerate(training_data['SMS']):
    for word in sms:
        word_counts_per_sms[word][index] += 1
```

In [14]:

```python
# Print N number of items in a dictionary

import itertools

n = 1
dict_name = word_counts_per_sms
out1 = dict(itertools.islice(dict_name.items(), n))
print(out1)
```

```
{'silent': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
```

0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,

```
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]}
```

For key word, "river," it looks like it has just appeared in one SMS message.

In [15]:

```python
# Convert dictionary, word_counts_per_sms, to dataframe

word_counts = pd.DataFrame(word_counts_per_sms)
word_counts.head(3)
```

Out[15]:

	silent	luv	offcampus	omg	joking	treat	charlie	86688	vijay	jez	...	gift	5pm	wtlp	budget	gain	aspects	goss
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0

3 rows × 7783 columns

The newly created dictionary, word_counts_per_sms, of unique vocabulary words and word counts, was converted to a dataframe, word_counts. Now the dataframe can be combined with the training dataset.

In [16]:

```python
# Combine training dataset with word_counts

training = pd.concat([training_data, word_counts], axis=1)
print(training.shape)
training.head(3)
```

```
(4458, 7785)
```

Out[16]:

	Label	SMS	silent	luv	offcampus	omg	joking	treat	charlie	86688	...	gift	5pm	wtlp	budget	gain	aspec
0	ham	[yep, by, the, pretty, sculpture]	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
1	ham	[yes, princess, are, you, going, to, make, me,...	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
2	ham	[welp, apparently, he, retired]	0	0	0	0	0	0	0	0	...	0	0	0	0	0	

3 rows × 7785 columns

Classify messages as spam or nonspam using Multinomial Naive Bayes Algorithm

First, using the cleaned and transformed training dataset, constants and parameters will be computed. Calculating these values before classification of messages makes Naive Bayes algorithm faster.

The following constants will be computed:

- probability of a spam message or nonspam message
- number of words in all spam messages
- number of words in all nonspam messages
- number of words in vocabulary list

The parameters are P(word|Spam) and P(word|nonSpam). Dictionaries of the parameters will be created.

Then the spam filter will be created.

Calculate Constants

The Naive Bayes algorithm will need to answer these two probability questions to be able to classify new messages, where w=word:

$$P(Spam|w_1, w_2, \ldots, w_n) \propto P(Spam) \cdot \prod_{i=1}^{n} P(w_i|Spam)$$

$$P(nonSpam|w_1, w_2, \ldots, w_n) \propto P(nonSpam) \cdot \prod_{i=1}^{n} P(w_i|nonSpam)$$

Calculate $P(w_i|Spam)$ and $P(w_i|nonSpam)$ using these equations:

$$P(w_i|Spam) = \frac{N_{w_i|Spam} + \alpha}{N_{Spam} + \alpha \cdot N_{Vocabulary}}$$

$$P(w_i|nonSpam) = \frac{N_{w_i|nonSpam} + \alpha}{N_{nonSpam} + \alpha \cdot N_{Vocabulary}}$$

The following terms will be computed and used repeatedly to classify messages:

- **P(Spam) and P(nonSpam)**
- **N_{Spam}, $N_{nonSpam}$, $N_{Vocabulary}$**

We'll also use Laplace smoothing and set $alpha = 1$.

In [17]:

```
# For the Training set, constants will be computed

spam_messages = training[training['Label'] == 'spam']
nonspam_messages = training[training['Label'] == 'ham']

# Probability of spam or nonspam message, p_spam & p_nonspam
p_spam = len(spam_messages) / len(training)
p_nonspam = len(nonspam_messages) / len(training)

# Number of words in all spam messages, n_spam
n_words_spam = spam_messages['SMS'].apply(len)
n_spam = n_words_spam.sum()

# Number of words in all nonspam messages, n_nonspam
n_words_nonspam = nonspam_messages['SMS'].apply(len)
n_nonspam = n_words_nonspam.sum()

# Number of vocabulary words, n_vocab
n_vocab = len(vocabulary)

# Laplace smoothing
alpha = 1

print(p_spam, p_nonspam)
print(n_spam, n_nonspam)
```

```
0.13458950201884254 0.8654104979811574
15190 57237
```

Calculate Parameters

Next, the parameters, $P(w_i|Spam)$ and $P(w_i|nonSpam)$, will be calculated.

The parameters are calculated using the formulas:

$$P(w_i|Spam) = \frac{N_{w_i|Spam} + \alpha}{N_{Spam} + \alpha}$$

$$\cdot N_{Vocabulary}$$

$$P(w_i|nonSpam) = \frac{N_{w_i|nonSpam} + \alpha}{N_{nonSpam} + \alpha \cdot N_{Vocabulary}}$$

In [18]:

```
# Create 2 dictionaries for the parameters P(word|Spam) and P(word|nonSpam)
# Keys are unique words from vocabulary list;
# values are the probability that the word given spam, or word given nonspam

# Initiate parameters in dictionaries: keys=word from vocabulary; value=0
spam_parameters = {unique_word:0 for unique_word in vocabulary}
nonspam_parameters = {unique_word:0 for unique_word in vocabulary}

for word in vocabulary:
    n_word_given_spam = spam_messages[word].sum()
    p_word_given_spam = (n_word_given_spam + alpha) / (
        n_spam + alpha*n_vocab)
    spam_parameters[word] = p_word_given_spam

    n_word_given_nonspam = nonspam_messages[word].sum()
    p_word_given_nonspam = (n_word_given_nonspam + alpha) / (
        n_nonspam + alpha*n_vocab)
    nonspam_parameters[word] = p_word_given_nonspam
```

In [19]:

```
# Print N items in dictionaries

import itertools

n = 10
dict_name = spam_parameters
# dictionary = nonspam_parameters
out2 = dict(itertools.islice(dict_name.items(), n))
print(out2)
```

```
{'silent': 4.3529360553693465e-05, 'luv': 0.00021764680276846734, 'offcampus': 4.3529
360553693465e-05, 'omg': 4.3529360553693465e-05, 'joking': 4.3529360553693465e-05, 't
reat': 4.3529360553693465e-05, 'charlie': 4.3529360553693465e-05, '86688': 0.00073999
91294127889, 'vijay': 4.3529360553693465e-05, 'jez': 4.3529360553693465e-05}
```

Classify a New Message

For an SMS message, the probability of it being spam or nonspam is given by:

$$P(Spam|w_1, w_2, \ldots, w_n) \propto P(Spam)$$

A function, classify, will be created that:

- **As input, takes in an SMS message ($w_1, w_2, \ldots, w_n$)**
- **calculates P(Spam|word1, word2..) & P(nonSpam|word1...)**
- **Classifies the SMS message as spam or nonspam based on:**
    - **if P(nonSpam|$w_1, w_2, \ldots, w_n$) > P(Spam|$w_1, w_2, \ldots, w_n$), then the message is classified as "ham" or nonspam.**
    - **if P(nonSpam|$w_1, w_2, \ldots, w_n$) < P(Spam|$w_1, w_2, \ldots, w_n$), then the message is classified as spam.**

- If P(nonSpam|$w_1, w_2, \ldots, w_n$) = P(Spam|$w_1, w_2, \ldots, w_n$), then the function returns a message requesting human classification

The spam filter can be understood as a function that:

- **Takes in as input a new message (w_1, w_2, ..., w_n).**
- **Compares the values of P(Spam|w_1, w_2, ..., w_n) and P(Ham|w_1, w_2, ..., w_n)**
  - **If P(Ham|w_1, w_2, ..., w_n) > P(Spam|w_1, w_2, ..., w_n), then the message is classified as ham.**
  - **If P(Ham|w_1, w_2, ..., w_n) < P(Spam|w_1, w_2, ..., w_n), then the message is classified as spam.**
  - **If P(Ham|w_1, w_2, ..., w_n) = P(Spam|w_1, w_2, ..., w_n), then the algorithm may request human help.**

In [20]:

```
# Function, classify, to classify SMS messages as spam or nonspam

import re

def classify(message):
    '''message is a string type'''
    message = re.sub('\W', ' ', message) #removes punctuation
    message = message.lower().split()

    p_spam_given_message = p_spam
    p_nonspam_given_message = p_nonspam

    for word in message:
        if word in spam_parameters:
            p_spam_given_message *= spam_parameters[word]
        if word in nonspam_parameters:
            p_nonspam_given_message *= nonspam_parameters[word]

    print('P(Spam|message): ', p_spam_given_message)
    print('P(nonSpam|message): ', p_nonspam_given_message)

    if p_nonspam_given_message > p_spam_given_message:
         print('Label: Not Spam')
    if p_nonspam_given_message < p_spam_given_message:
        print('Label: Spam')
    else:
        print('Equal probabilities, human needs to classify message.')
```

In [21]:

```
# Test classify function.  message1 should be 'Not Spam'
message1 = 'Sounds good, Tom, then see u there'
print(message1, classify(message1))
```

```
P(Spam|message):  2.4372375665888117e-25
P(nonSpam|message):  3.687530435009238e-21
Label: Not Spam
Equal probabilities, human needs to classify message.
Sounds good, Tom, then see u there None
```

In [22]:

```
# Test classify function.  message2 should be 'Spam'
message2 = 'WINNER!! This is the secret code to unlock the money: C3421.'
print(message2, classify(message2))
```

```
P(Spam|message):  1.3481290211300841e-25
P(nonSpam|message):  1.9368049028589875e-27
Label: Spam
WINNER!! This is the secret code to unlock the money: C3421. None
```

Apply Classification on Testing dataset

Testing the classify function on the testing dataset will provide a measure of accuracy of the function.

In [23]:

```
# Function to classify Testing data set: returns classifications rather than prints t
hem

import re

def classify_test_data(message):
    message = re.sub('\W', ' ', message) #removes punctuation
    message = message.lower().split()

    p_spam_given_message = p_spam
    p_nonspam_given_message = p_nonspam

    for word in message:
        if word in spam_parameters:
            p_spam_given_message *= spam_parameters[word]
        if word in nonspam_parameters:
            p_nonspam_given_message *= nonspam_parameters[word]

    if p_nonspam_given_message > p_spam_given_message:
        return 'ham'
    if p_nonspam_given_message < p_spam_given_message:
        return 'spam'
    else:
        return 'requires human classification'
```

In [24]:

```
'''To avoid a SettingWithCopy Warning'''
testing = testing.copy()

testing['predicted'] = testing['SMS'].apply(classify_test_data)
testing.head()
```

Out[24]:

	Label	SMS	predicted
0	ham	Later i guess. I needa do mcat study too.	ham
1	ham	But i haf enuff space got like 4 mb...	ham
2	spam	Had your mobile 10 mths? Update to latest Oran...	spam
3	ham	All sounds good. Fingers . Makes it difficult ...	ham
4	ham	All done, all handed in. Don't know if mega sh...	ham

Accuracy of Spam Filter

In [25]:

```
correct = 0
total = len(testing)

for row in testing.iterrows():
    row = row[1]
    if row['Label'] == row['predicted']:
        correct += 1
    accuracy = correct/total
accuracy
```

Out[25]:

```
0.9874326750448833
```

The spam filter is 98.7% accurate, which exceeds our goal of 80% accuracy! The spam filter looked at 1,114 SMS messages, and classified 1,100 correctly. In other words, the spam filter is highly reliable.

CONCLUSION

The goal of this project was to create a spam filter that classifies messages as spam or not spam. This was accomplished using the multinomial Naive Bayes algorithm, which is based on conditional probability.

The dataset was split into two datasets, a training dataset, and a testing dataset. The filter was applied to the training dataset, and then tested using the testing dataset. It was found that the spam filter is 98.7% accurate, which exceeded the initial goal of 80% accuracy.

For future iterations of this project, possible next steps are:

- Determine why the algorithm misclassified the 14 messages.
- The algorithm could be made more complex such as making the algorithm sensitive to letter case.