Creado por:

Isabel Maniega

Natural Language Processing (NLP)

https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

El objetivo de este ejercicio:

- Los ordenadores trabajan con números, no con letras
- así que necesitamos NLP para tranasformar las palabras a números

---

In [1]:
```python
import warnings
warnings.filterwarnings("ignore")
```

In [2]:
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [4]:
```python
from sklearn.naive_bayes import MultinomialNB
```

Cargar archivo .csv

In [5]:
```python
# https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset
```

In [6]:
```python
df = pd.read_csv("spam.csv",
                 sep=",", encoding='ISO-8859-1')
df.head(15)
```

Out[6]:

	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
5	spam	FreeMsg Hey there darling it's been 3 week's n...	NaN	NaN	NaN
6	ham	Even my brother is not like to speak with me. ...	NaN	NaN	NaN
7	ham	As per your request 'Melle Melle (Oru Minnamin...	NaN	NaN	NaN
8	spam	WINNER!! As a valued network customer you have...	NaN	NaN	NaN
9	spam	Had your mobile 11 months or more? U R entitle...	NaN	NaN	NaN
10	ham	I'm gonna be home soon and i don't want to tal...	NaN	NaN	NaN
11	spam	SIX chances to win CASH! From 100 to 20,000 po...	NaN	NaN	NaN
12	spam	URGENT! You have won a 1 week FREE membership ...	NaN	NaN	NaN
13	ham	I've been searching for the right words to tha...	NaN	NaN	NaN
14	ham	I HAVE A DATE ON SUNDAY WITH WILL!!	NaN	NaN	NaN

In [7]:
```python
df = df.iloc[:, 0:2]
df.head()
```

Out[7]:

	v1	v2
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...

Nombres para las columnas

In [8]:
```python
df.columns= ["Status", "Message"]
df.head()
```

Out[8]:

	Status	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...

In [10]:
```python
df.shape
```

Out[10]:
```
(5572, 2)
```

In [11]:
```python
len(df)
```

Out[11]:
```
5572
```

Vemos si nos faltan algunos datos

In [12]:
```python
df.Message.isnull().sum()
```

Out[12]:
```
0
```

In [13]:
```python
df.describe()
```

Out[13]:

	Status	Message
count	5572	5572
unique	2	5169
top	ham	Sorry, I'll call later
freq	4825	30

¿Cuántos datos de "spam" en nuestros datos?

Forma 1

In [14]:
```python
df.head()
```

Out[14]:

	Status	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...

In [15]:
```python
df.Status.value_counts()
```

Out[15]:
```
ham     4825
spam     747
Name: Status, dtype: int64
```

Forma 2

In [16]:
```python
df.iloc[:,0].value_counts()
```

Out[16]:
```
ham     4825
spam     747
Name: Status, dtype: int64
```

Forma 3

In [17]:
```python
df_spam = df[df.Status == "spam"]
len(df_spam)
```

Out[17]:
```
747
```

Forma 4

In [18]:
```python
data = df[df.iloc[:,0] == "spam"]
len(data)
```

Out[18]:
```
747
```

spam == 1 (True); ham == 0 (False)

Método 1

In [19]:
```python
df["Status"] = df["Status"].map({"ham": 0, "spam": 1})
df.head()
```

Out[19]:

	Status	Message
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]:
```python
df.shape
```

Out[20]:
```
(5572, 2)
```

In [21]:
```python
X = df.Message
```

In [22]:
```python
y = df.Status
```

Train, Test split

In [24]:
```python
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0, test_size=0.2)
```

Método 1: CountVectorizer

In [25]:
```python
from sklearn.feature_extraction.text import CountVectorizer
```

In [26]:
```python
cv = CountVectorizer()
```

In [28]:
```python
X_train = cv.fit_transform(X_train)
X_test = cv.transform(X_test)
```

In [29]:
```python
y_train = y_train.astype("int")
y_test = y_test.astype("int")
```

In [30]:
```python
y_train = np.array(y_train)
y_test = np.array(y_test)
```

In [31]:
```python
X_train
```

Out[31]:
```
<4457x7612 sparse matrix of type '<class 'numpy.int64'>'
	with 58826 stored elements in Compressed Sparse Row format>
```

In [32]:
```python
X_test
```

Out[32]:
```
<1115x7612 sparse matrix of type '<class 'numpy.int64'>'
	with 13975 stored elements in Compressed Sparse Row format>
```

In [33]:
```python
y_train
```

Out[33]:
```
array([0, 0, 0, ..., 0, 0, 0])
```

In [34]:
```python
y_test
```

Out[34]:
```
array([0, 0, 0, ..., 0, 0, 0])
```

Un poco de Machine Learning

In [35]:
```python
clf = MultinomialNB()
```

In [36]:
```python
clf.fit(X_train, y_train)
```

Out[36]:
```
▾ MultinomialNB
MultinomialNB()
```

In [38]:
```python
y_pred = clf.predict(X_test)
y_pred
```

Out[38]:
```
array([0, 0, 0, ..., 0, 0, 0])
```

In [39]:
```python
from sklearn.metrics import accuracy_score
acc = accuracy_score(y_pred, y_test)
print(acc * 100)
```

```
98.7443946188341
```

In [40]:
```python
clf.score(X_test, y_test)
```

Out[40]:
```
0.9874439461883409
```

In [42]:
```python
aciertos = 0

for i in range(len(y_pred)):
    if y_pred[i] == y_test[i]:
        aciertos += 1
aciertos
```

Out[42]:
```
1101
```

In [43]:
```python
(aciertos/len(y_pred))*100
```

Out[43]:
```
98.7443946188341
```

Calcular la matriz de confusión

In [44]:
```python
len(y_train)
```

Out[44]:
```
4457
```

Falsos Positivos

In [46]:
```python
FP = 0

for i in np.arange(len(y_test)):
    if y_test[i] == 0 and y_pred[i] == 1:
        FP += 1
FP
```

Out[46]:
```
2
```

Falsos Negativos

In [47]:
```python
FN = 0

for i in np.arange(len(y_test)):
    if y_test[i] == 1 and y_pred[i] == 0:
        FN += 1
FN
```

Out[47]:
```
12
```

True Positives

In [48]:
```python
TP = 0

for i in np.arange(len(y_test)):
    if y_test[i] == 1 and y_pred[i] == 1:
        TP += 1
TP
```

Out[48]:
```
154
```

True Negative

In [49]:
```python
TN = 0

for i in np.arange(len(y_test)):
    if y_test[i] == 0 and y_pred[i] == 0:
        TN += 1
TN
```

Out[49]:
```
947
```

In [51]:
```python
confusion_matrix = np.array([[TN, FP],
                             [FN, TP]])
confusion_matrix
```

Out[51]:
```
array([[947,   2],
       [ 12, 154]])
```

In [52]:
```python
((TN + TP) / (TN+TP+FP+FN)) *100
```

Out[52]:
```
98.7443946188341
```

Forma con Sklearn

In [53]:
```python
from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test, y_pred)
cm
```

Out[53]:
```
array([[947,   2],
       [ 12, 154]])
```

Ahora con: TfidfVectorizer

In [54]:
```python
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0, test_size=0.2)
```

In [55]:
```python
X_train
```

Out[55]:
```
1114    No no:)this is kallis home ground.amla home to...
3589    I am in escape theatre now. . Going to watch K...
3095    We walked from my moms. Right on stagwood pass...
1012       I dunno they close oredi not... ÌÏ v ma fan...
3320                               Yo im right by yo work
                              ...
4931                Match started.india  &lt;#&gt;  for 2
3264    44 7732584351, Do you want a New Nokia 3510i c...
1653    I was at bugis juz now wat... But now i'm walk...
2607    :-) yeah! Lol. Luckily i didn't have a starrin...
2732    How dare you stupid. I wont tell anything to y...
Name: Message, Length: 4457, dtype: object
```

In [56]:
```python
X_test
```

Out[56]:
```
4456    Aight should I just plan to come up later toni...
690                                Was the farm open?
944     I sent my scores to sophas and i had to do sec...
3768    Was gr8 to see that message. So when r u leavi...
1189    In that case I guess I'll see you at campus lodge
                              ...
2906                                           ALRITE
1270    Sorry chikku, my cell got some problem thts y ...
3944    I will be gentle princess! We will make sweet ...
2124    Beautiful Truth against Gravity.. Read careful...
253     Ups which is 3days also, and the shipping comp...
Name: Message, Length: 1115, dtype: object
```

In [57]:
```python
y_train
```

Out[57]:
```
1114    0
3589    0
3095    0
1012    0
3320    0
       ..
4931    0
3264    1
1653    0
2607    0
2732    0
Name: Status, Length: 4457, dtype: int64
```

In [58]:
```python
y_test
```

Out[58]:
```
4456    0
690     0
944     0
3768    0
1189    0
       ..
2906    0
1270    0
3944    0
2124    0
253     0
Name: Status, Length: 1115, dtype: int64
```

In [59]:
```python
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [60]:
```python
tv = TfidfVectorizer(stop_words = "english")
tv
```

Out[60]:
```
▾          TfidfVectorizer
TfidfVectorizer(stop_words='english')
```

In [61]:
```python
X_train = tv.fit_transform(X_train)
X_test = tv.transform(X_test)
```

In [62]:
```python
y_train = y_train.astype("int")
y_test = y_test.astype("int")
```

In [63]:
```python
y_train = np.array(y_train)
y_test = np.array(y_test)
```

Creamos el algoritmo

In [64]:
```python
clf = MultinomialNB()
```

In [65]:
```python
clf.fit(X_train, y_train)
```

Out[65]:
```
▾ MultinomialNB
MultinomialNB()
```

In [66]:
```python
y_pred = clf.predict(X_test)
y_pred
```

Out[66]:
```
array([0, 0, 0, ..., 0, 0, 0])
```

In [67]:
```python
from sklearn.metrics import accuracy_score
acc = accuracy_score(y_pred, y_test)
print(acc * 100)
```

```
96.59192825112108
```

In [68]:
```python
from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test, y_pred)
cm
```

Out[68]:
```
array([[949,   0],
       [ 38, 128]])
```

Creado por:

Isabel Maniega