In [4]:
```
import pandas as pd
import numpy as np

data = pd.read_csv("language.csv")
data
```

Out[4]:

	Text	language
0	klement gottwaldi surnukeha palsameeriti ning ...	Estonian
1	sebes joseph pereira thomas på eng the jesuit...	Swedish
2	ถนนเจริญกรุง อักษรโรมัน thanon charoen krung เ...	Thai
3	விசாகப்பட்டினம் தமிழ்ச்சங்கத்தை இந்துப் பத்திர...	Tamil
4	de spons behoort tot het geslacht haliclona en...	Dutch
...	...	...
21995	hors du terrain les années et sont des année...	French
21996	ใน พศ หลักจากที่เสด็จประพาสแหลมมลายู ชวา อินเ...	Thai
21997	con motivo de la celebración del septuagésimoq...	Spanish
21998	年月，當時還只有歲的她在美國出道，以mai-k名義推出首張英文《baby i like》，由...	Chinese
21999	aprilie sonda spațială messenger a nasa și-a ...	Romanian

22000 rows × 2 columns

In [5]:
```
from sklearn.feature_extraction.text import CountVectorizer
```

In [6]:
```
from sklearn.model_selection import train_test_split
```

In [7]:
```
from sklearn.naive_bayes import MultinomialNB
```

In [8]:
```
data
```

Out[8]:

	Text	language
0	klement gottwaldi surnukeha palsameeriti ning ...	Estonian
1	sebes joseph pereira thomas på eng the jesuit...	Swedish
2	ถนนเจริญกรุง อักษรโรมัน thanon charoen krung เ...	Thai
3	விசாகப்பட்டினம் தமிழ்ச்சங்கத்தை இந்துப் பத்திர...	Tamil
4	de spons behoort tot het geslacht haliclona en...	Dutch
...	...	...
21995	hors du terrain les années et sont des année...	French
21996	ใน พศ หลักจากที่เสด็จประพาสแหลมมลายู ชวา อินเ...	Thai
21997	con motivo de la celebración del septuagésimoq...	Spanish
21998	年月，當時還只有歲的她在美國出道，以mai-k名義推出首張英文《baby i like》，由...	Chinese
21999	aprilie sonda spațială messenger a nasa și-a ...	Romanian

22000 rows × 2 columns

In [9]:
```
data.isnull().sum()
```

Out[9]:
```
Text        0
language    0
dtype: int64
```

In [10]:
```
data['language'].value_counts()
```

Out[10]:
```
language
Estonian      1000
Swedish       1000
English       1000
Russian       1000
Romanian      1000
Persian       1000
Pushto        1000
Spanish       1000
Hindi         1000
Korean        1000
Chinese       1000
French        1000
Portugese     1000
Indonesian    1000
Urdu          1000
Latin         1000
Turkish       1000
Japanese      1000
Dutch         1000
Tamil         1000
Thai          1000
Arabic        1000
Name: count, dtype: int64
```

In [11]:
```
data.dtypes
```

Out[11]:
```
Text        object
language    object
dtype: object
```

In [12]:
```
X = np.array(data['Text'])
y = np.array(data['language'])
```

In [13]:
```
print(X)
```

```
['klement gottwaldi surnukeha palsameeriti ning paigutati mausoleumi surnukeha ol
i aga liiga hilja ja oskamatult palsameeritud ning hakkas ilmutama lagunemise tun
demärke  aastal viidi ta surnukeha mausoleumist ära ja kremeeriti zlíni linn kand
is aastatel – nime gottwaldov ukrainas harkivi oblastis kandis zmiivi linn aastat
el – nime gotvald'
 'sebes joseph pereira thomas  på eng the jesuits and the sino-russian treaty of
nerchinsk  the diary of thomas pereira bibliotheca instituti historici s i --   r
ome libris '
 'ถนนเจริญกรุง อักษรโรมัน thanon charoen krung เริ่มตั้งแต่ถนนสนามไชยถึงแม่น้ำเจ้าพระยาที่ถนนตก
กรุงเทพมหานคร เป็นถนนรุ่นแรกที่ใช้เทคนิคการสร้างแบบตะวันตก ปัจจุบันผ่านพื้นที่เขตพระนคร เขตป้อมปรา
บศัตรูพ่าย เขตสัมพันธวงศ์ เขตบางรัก เขตสาทร และเขตบางคอแหลม'
 ...
 'con motivo de la celebración del septuagésimoquinto º aniversario de la fundaci
ón del departamento en  guillermo ceballos espinosa presentó a la gobernación de
caldas por encargo de su titular dilia estrada de gómez el himno que fue adoptado
para solemnizar dicha efemérides y que siguieron interpretando las bandas de músi
ca y los planteles de educación de esta sección del país en retretas y actos ofic
iales con gran aceptación[]\u200b'
 '年月，當時還只有歲的她在美國出道，以mai-k名義推出首張英文《baby i like》，由美國的獨
立廠牌bip·record發行，以外國輸入盤的形式在日本發售，旋即被抢购一空。其後於月日發行以倉木
麻衣名義發行的首張日文單曲《love day after tomorrow》，正式於日本出道。這張單曲初動銷量
只得約萬張，可是其後每週銷量一直上升，並於年月正式突破百萬銷量，合计万张。成為年最耀眼的
新人歌手。'
 ' aprilie sonda spațială messenger a nasa și-a încheiat misiunea de studiu de  a
ni prăbușindu-se pe suprafața planetei mercur sonda a rămas fără combustibil fiin
d împinsă de gravitația solară din ce în ce mai aproape de mercur']
```

In [14]:
```
print(y)
```

```
['Estonian' 'Swedish' 'Thai' ... 'Spanish' 'Chinese' 'Romanian']
```

In [15]:
```
cv = CountVectorizer()
X = cv.fit_transform(X)
```

In [16]:
```
X_train,X_test, y_train,y_test = train_test_split(X,y, test_size = 0.33,random_s
```

In [17]:
```
X_train
```

Out[17]:
```
<14740x277720 sparse matrix of type '<class 'numpy.int64'>'
        with 613529 stored elements in Compressed Sparse Row format>
```

In [18]:
```
print(X_train)
```

```
(0, 197295)   2
(0, 197708)   1
(0, 197801)   1
(0, 198388)   1
(0, 197467)   1
(0, 197865)   2
(0, 197604)   1
(0, 198428)   1
(0, 198501)   1
(0, 198556)   1
(0, 197332)   1
(0, 197485)   2
(0, 198123)   1
(0, 197892)   1
(0, 197990)   1
(0, 198053)   1
(0, 198417)   1
(0, 197623)   1
(1, 197641)   2
(1, 197314)   1
(1, 197931)   1
(1, 197804)   3
(1, 198397)   1
(1, 197149)   1
(1, 197781)   1
:     :
(14738, 188817)       1
(14738, 192004)       1
(14738, 157171)       1
(14738, 190346)       1
(14738, 190725)       1
(14738, 189685)       1
(14738, 159269)       2
(14738, 145431)       1
(14738, 173292)       1
(14738, 176062)       1
(14738, 159959)       1
(14738, 190198)       1
(14738, 167124)       1
(14738, 168158)       1
(14738, 180260)       2
(14738, 153262)       1
(14738, 162150)       1
(14738, 153355)       1
(14738, 178104)       1
(14738, 163770)       1
(14739, 223002)       1
(14739, 235170)       1
(14739, 222446)       1
(14739, 221922)       1
(14739, 242446)       1
```

In [19]:
```
print(y_test)
```

```
['Japanese' 'Russian' 'Latin' ... 'Turkish' 'Arabic' 'English']
```

In [20]:
```
model = MultinomialNB()
```

In [21]:
```
model.fit(X_train,y_train)
```

Out[21]:
```
MultinomialNB
MultinomialNB()
```

In [22]:
```
model.score(X_test,y_test)
```

Out[22]:
```
0.953168044077135
```

In [23]:
```
user = input("Enter a text")
data = cv.transform([user]).toarray()
output = model.predict(data)
print(output)
```

```
['English']
```

In [ ]:
```
```