Language Detection

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

In [2]:

```
df=pd.read_csv("D://Language Detection.csv")
df.head(10)
```

Out[2]:

	Text	Language
0	Nature, in the broadest sense, is the natural...	English
1	"Nature" can refer to the phenomena of the phy...	English
2	The study of nature is a large, if not the onl...	English
3	Although humans are part of nature, human acti...	English
4	[1] The word nature is borrowed from the Old F...	English
5	[2] In ancient philosophy, natura is mostly us...	English
6	[3][4] \nThe concept of nature as a whole, the...	English
7	During the advent of modern scientific method ...	English
8	[5][6] With the Industrial revolution, nature ...	English
9	However, a vitalist vision of nature, closer t...	English

In [3]:

```
df.shape
```

Out[3]:

```
(10337, 2)
```

In [4]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10337 entries, 0 to 10336
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Text      10337 non-null  object
 1   Language  10337 non-null  object
dtypes: object(2)
memory usage: 161.6+ KB
```

In [5]:
```
df.isna().sum()
```

Out[5]:
```
Text        0
Language    0
dtype: int64
```

In [6]:
```
df['Language'].value_counts()
```

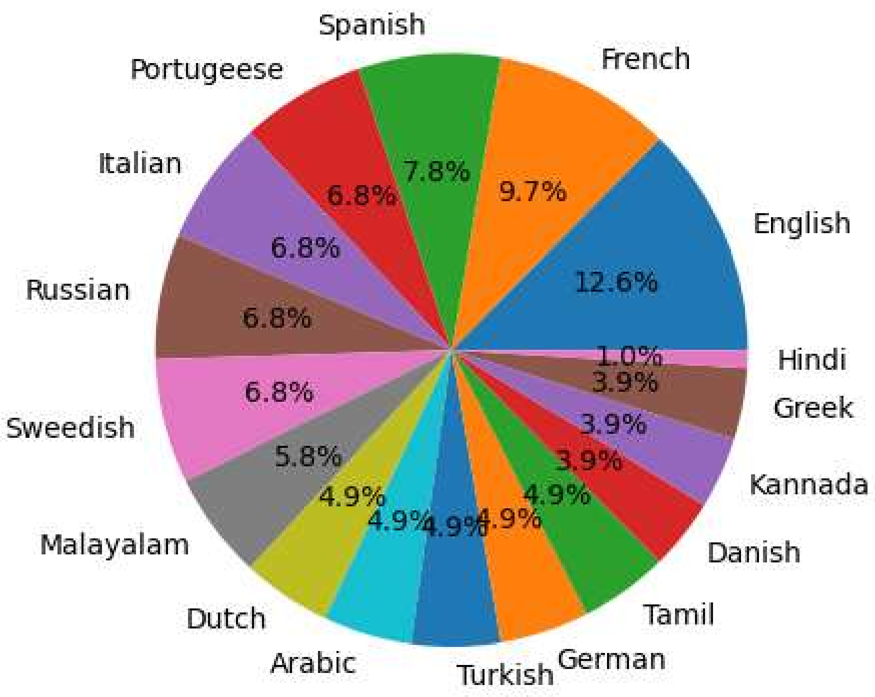
Out[6]:
```
English       1385
French        1014
Spanish        819
Portugeese     739
Italian        698
Russian        692
Sweedish       676
Malayalam      594
Dutch          546
Arabic         536
Turkish        474
German         470
Tamil          469
Danish         428
Kannada        369
Greek          365
Hindi           63
Name: Language, dtype: int64
```

In our dataset there are 17 different language are their

In [7]:
```
labels=['English','French','Spanish','Portugeese','Italian','Russian','Sweedis
       'Tamil','Danish','Kannada','Greek','Hindi']
```

In [8]:
```
data=np.round(df['Language'].value_counts()/df.shape[0]*100)
```

In [9]:
```
plt.pie(data,labels=labels,autopct='%1.1f%%')
plt.show()
```



In [10]:
```
x = np.array(df["Text"])
y = np.array(df["Language"])
```

In [11]:
```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
```

In [12]:
```
cv = CountVectorizer()
X = cv.fit_transform(x)
X_train, X_test, y_train, y_test = train_test_split(X, y,test_size=0.33,random
```

In [13]:
```
model = MultinomialNB()
model.fit(X_train,y_train)
np.round(model.score(X_test,y_test)*100)
```

Out[13]:
```
98.0
```

accuracy is approx 98% that is good for the model

In [14]:

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

```
Enter a Text: हिन्दी जिसके मानकीकृत रूप को मानक हिन्दी कहा जाता है, विश्व की एक प्रमुख
भाषा है एवं भारत की एक राजभाषा है। केन्द्रीय स्तर पर भारत में सह-आधिकारिक भाषा अंग्रेजी है।
यह हिन्दुस्तानी भाषा की एक मानकीकृत रूप है जिसमें संस्कृत के तत्सम तथा तद्भव शब्दों का प्रयोग
अधिक है और अरबी-फ़ारसी शब्द कम हैं।
['Hindi']
```

In [ ]:

```
```