ASSESSMENT FOR DATA SCIENCE TRAINEE

Web scraping is a skill I feel every data science enthusiast should know. It is immensely helpful when we're looking for data for our project or want to analyze specific data present only on a website. Keep in mind though, web scraping should not cross ethical and legal boundaries.

In this project, I use web scraping to extract YouTube video data using Google Developer and Python. We will then use the NLTK library to clean the data and then build a model to classify these videos based on specific categories.

Step 1 :

Install Jupyter

Step 2 :

Collect all synonymous word

Step 3 :

Visit at https://console.developers.google.com/

Extract data from youtube API and collect a jupyter notebook in csv format

Category of Data

- Travel
- Science
- Food
- History
- Manufacturing
- Art & Dance

Extract data like Id,Title,Description and merge all data in one DataFrame

Step 3 :

Remove all Duplicates in given data.

Step 4 :

In this section, we'll use the popular NLTK library to clean the data present in the "title" and "description" columns.

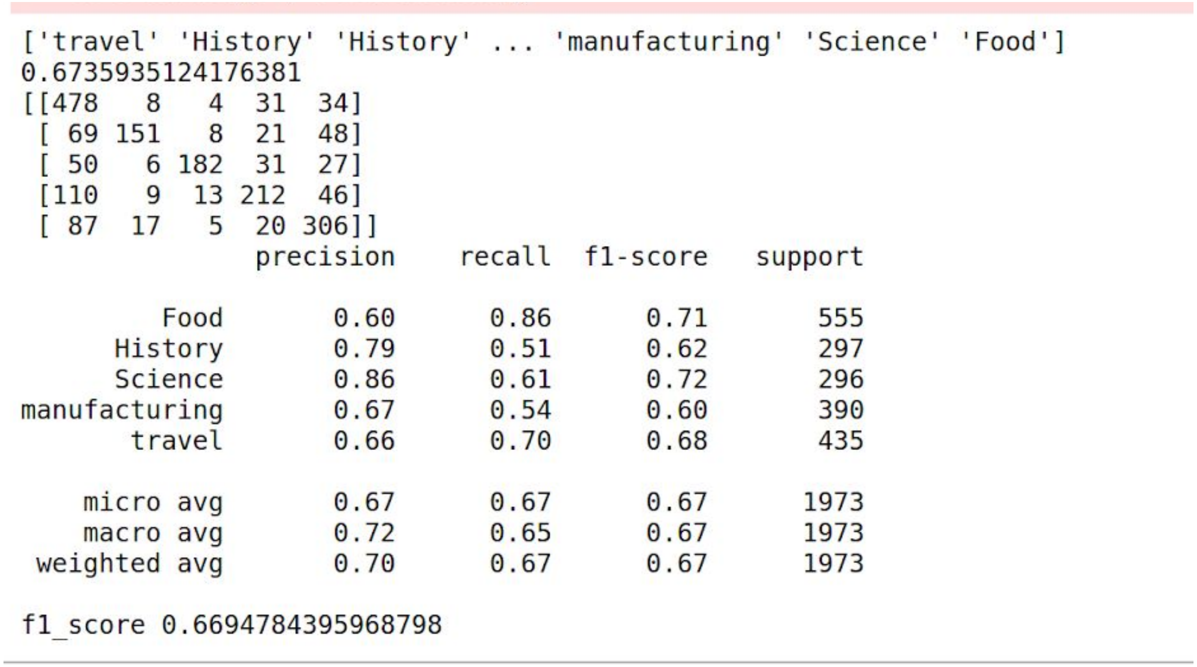
Before we start cleaning the data, we need to store all the columns separately so that we can perform different operations quickly and easily:

Step 5 :

Split data in train and test data format than Vectorize the all data using Tf-Idf

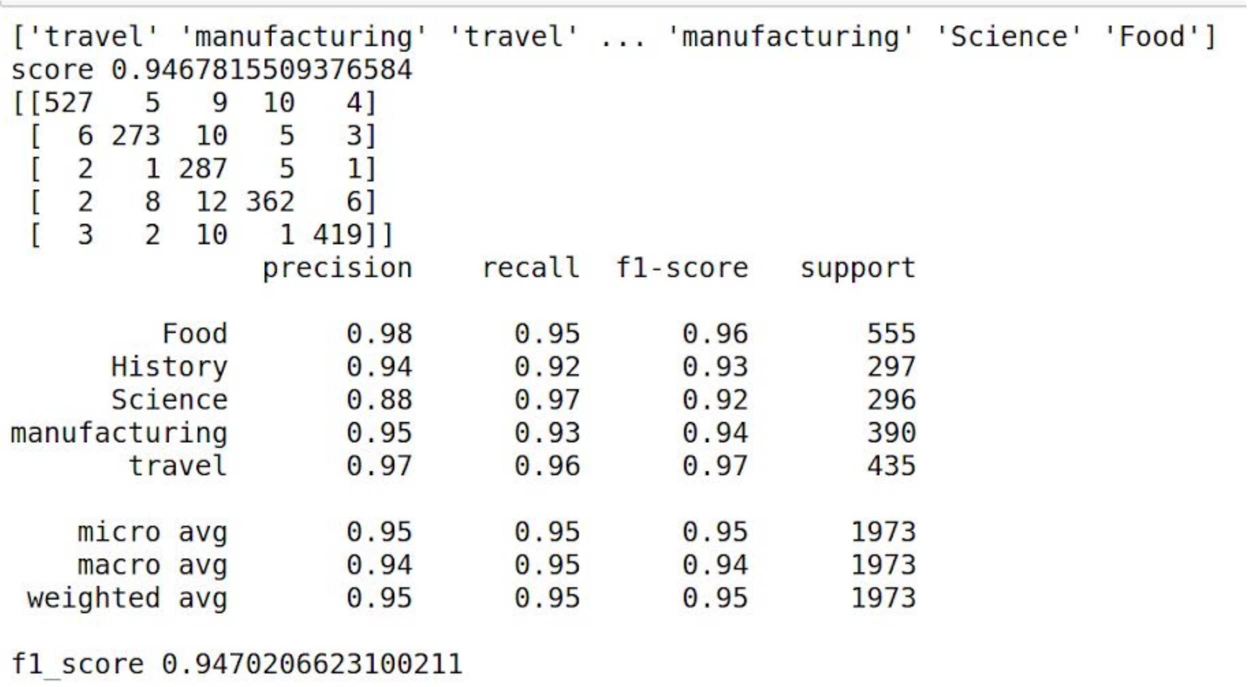
Step 6 :

From Categorie 1 I use Logistic Regression.



```
['travel' 'History' 'History' ... 'manufacturing' 'Science' 'Food']
0.6735935124176381
[[478   8   4  31  34]
 [ 69 151   8  21  48]
 [ 50   6 182  31  27]
 [110   9  13 212  46]
 [ 87  17   5  20 306]]
               precision    recall  f1-score   support

         Food       0.60      0.86      0.71       555
      History       0.79      0.51      0.62       297
      Science       0.86      0.61      0.72       296
manufacturing       0.67      0.54      0.60       390
       travel       0.66      0.70      0.68       435

    micro avg       0.67      0.67      0.67      1973
    macro avg       0.72      0.65      0.67      1973
 weighted avg       0.70      0.67      0.67      1973

f1_score 0.6694784395968798
```

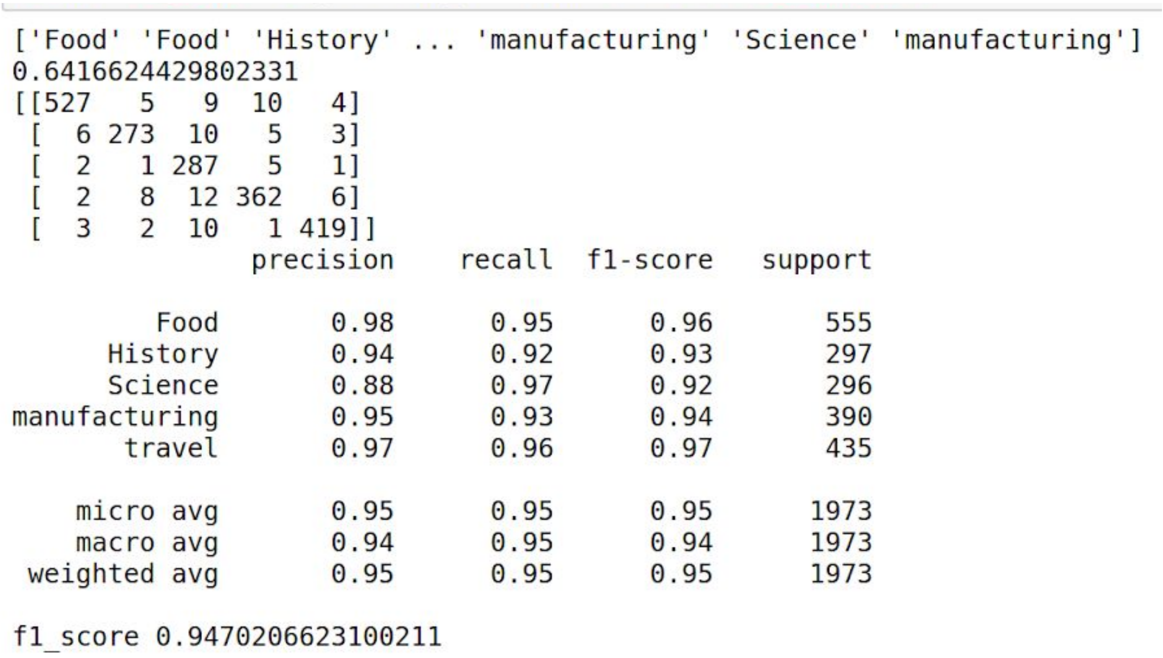
Naive-Bayes :



```
['travel' 'manufacturing' 'travel' ... 'manufacturing' 'Science' 'Food']
score 0.9467815509376584
[[527   5   9  10   4]
 [  6 273  10   5   3]
 [  2   1 287   5   1]
 [  2   8  12 362   6]
 [  3   2  10   1 419]]
               precision    recall  f1-score   support

         Food       0.98      0.95      0.96       555
      History       0.94      0.92      0.93       297
      Science       0.88      0.97      0.92       296
manufacturing       0.95      0.93      0.94       390
       travel       0.97      0.96      0.97       435

    micro avg       0.95      0.95      0.95      1973
    macro avg       0.94      0.95      0.94      1973
 weighted avg       0.95      0.95      0.95      1973

f1_score 0.9470206623100211
```

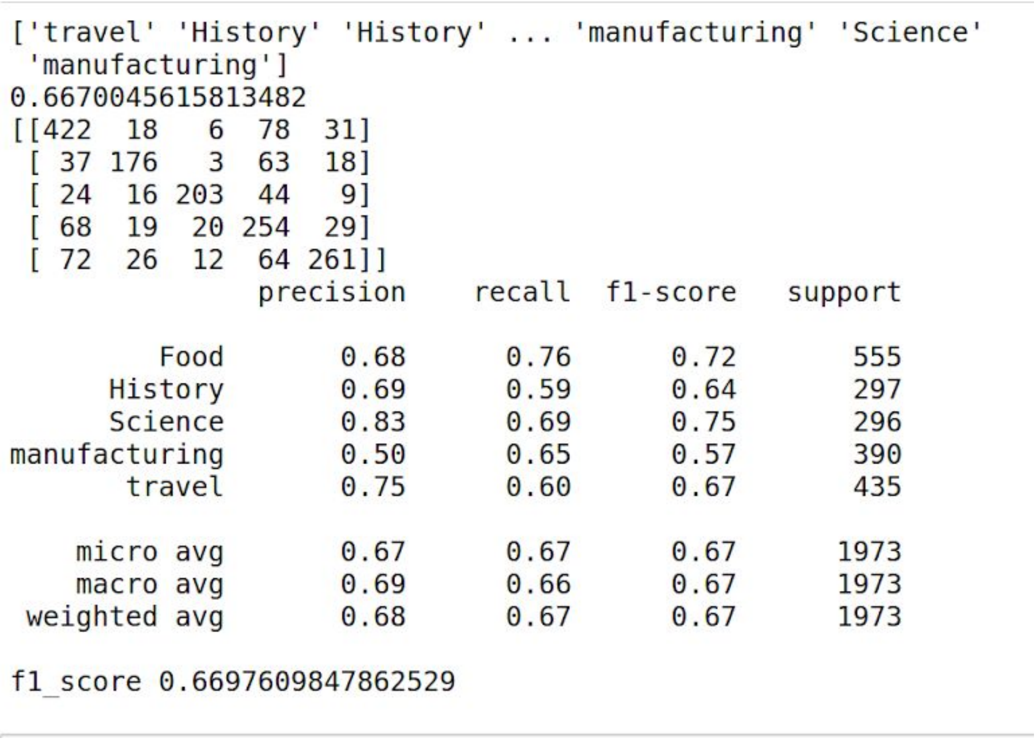
Step 7 :

From Categorie 2 I use Boosting.



```
['Food' 'Food' 'History' ... 'manufacturing' 'Science' 'manufacturing']
0.6416624429802331
[[527   5   9  10   4]
 [  6 273  10   5   3]
 [  2   1 287   5   1]
 [  2   8  12 362   6]
 [  3   2  10   1 419]]
               precision    recall  f1-score   support

         Food       0.98      0.95      0.96       555
      History       0.94      0.92      0.93       297
      Science       0.88      0.97      0.92       296
manufacturing       0.95      0.93      0.94       390
       travel       0.97      0.96      0.97       435

    micro avg       0.95      0.95      0.95      1973
    macro avg       0.94      0.95      0.94      1973
 weighted avg       0.95      0.95      0.95      1973

f1_score 0.9470206623100211
```

Bagging



```
['travel' 'History' 'History' ... 'manufacturing' 'Science'
 'manufacturing']
0.6670045615813482
[[422  18   6  78  31]
 [ 37 176   3  63  18]
 [ 24  16 203  44   9]
 [ 68  19  20 254  29]
 [ 72  26  12  64 261]]
               precision    recall  f1-score   support

         Food       0.68      0.76      0.72       555
      History       0.69      0.59      0.64       297
      Science       0.83      0.69      0.75       296
manufacturing       0.50      0.65      0.57       390
       travel       0.75      0.60      0.67       435

    micro avg       0.67      0.67      0.67      1973
    macro avg       0.69      0.66      0.67      1973
 weighted avg       0.68      0.67      0.67      1973

f1_score 0.6697609847862529
```