ASSESSMENT FOR DATA SCIENCE TRAINEE : REPORT

1. Summary on Scraping Techniques used.

   I used Youtube API key from Google developer Console (I had to create 6 keys from 6 separate projects, since the number of requests are limited for a single key/per day) and searched for the various topics such as Food, Travel Blogs , etc and got the description, title and video id. I generated 2000 instances for each class.

   I had initially tried to use selenium to automate the task of opening youtube in a browser, searching for the categories, and then opening each video from the categories (applied the filter : Videos only), then taking out full description, title, and video id and then going back to the search and going to the next video and thus, looping (would have to scroll the page, etc), but selenium crashed frequently, even when using headless option for the chromedriver (Good internet speed is required for even the headless version of chrome).

   I also wanted to learn the requests package and try using it to see if scraping can be done using it and beautiful soup, without using any APIs and thus, eliminating the restriction of number of requests and obtaining the full descriptions. I'll try to learn it and use it and see what can be achieved.

   In my github repo, you can find the following files:

   Youtube_scraper folder: youtube_scrape.py and youtube_videos.py # These are related to using the youtube api. I wrote youtube_scrape.py and youtube_videos.py is obtained from youtube's tutorials github.

   Selenium.py - It is the selenium web automater.

2. Models:

Model Types : Linear Classifiers, Naive Bayes Classifiers or SVMs.

I would use SVM as my classifier here. Reason against naive bayes is that for one thing, since my model is having balanced data ~2000 instances per category, the feature Pr(y) [y ⇒> target variable] is equal for all the classes. And, thus when finding the argument for a particular example naive bayes will only compare Pr(Xi| y) for different values of y, so number of parameters have lessened. Also naive bayes is used as a baseline/benchmark model to compare other models with.

SVMs tend to be the most accurate classifiers for text data, since it is high- dimensional. It also has strong theoretical foundations in optimization theory.

Why SVMs should be used:
https://stackoverflow.com/questions/35360081/naive-bayes-vs-svm-for-classifying-text-data

From scikit learn:

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

3. Precision, Recall and F1-score

These can be found as a dataframe table in the respective notebooks for the three different models also.

For SVM

Category	**Precision**	**Recall**	**F1_score**
Art and Music	**1**	**1**	**1**
Food	**1**	**1**	**1**
History	**1**	**0.998**	**0.998999**

Manufacturing	1	1	1
Science and Technology	0.997854	1	0.998926
Travel and Blogs	1	1	1

Model Types : Bagging models, boosting models or shallow NNs

I won’t use a bagging model such as random forest for text classification, since random forest is not a good choice for very high dimensional tasks compared to fast accurate linear models. I am using Xgboost due to its high performance and computation efficiency. Using shallow neural networks wouldn’t be useful since the number of neurons needed would be very high around 8000 for input and the single hidden layer may get overfit to the training data.

For XGBOOST

Category	Precision	Recall	F1_score
Art and Music	0.983015	1	0.991435
Food	0.998092	0.99619	0.99714
History	0.992016	0.994	0.993007
Manufacturing	1	0.997988	0.998993
Science and Technology	1	0.984946	0.992416
Travel and Blogs	1	1	1

Model Types : CNN, LSTM, GRU, Bidirectional RNNs, or RCNNS

4. EXPLANATION ABOUT THE RESULTS

Will do