ASSESSMENT FOR DATA SCIENCE TRAINEE

Scrap YouTube (or similar websites like Dailymotion) to collect data for classifying videos (based on description) in following 6 categories –

- Travel Blogs
- Science and Technology
- Food
- Manufacturing
- History
- Art and Music

For each of the aforementioned categories, scrape the video descriptions and prepare csv file which looks like the following:

Video id	Title	Description	Category
RcmrbNRK- jY	200 Days - A Trip Around the World Travel Film	My wife and I traveled to 17 countries in 200 days. This film is the story of our incredible trip! Enjoy! We used a GoPro and a Nikon D7000 for all of the filming. For Business Inquiries please e.mail me at 40northdesigns@gmail.com.	travel

Video id: It is the video URL which is unique for any video on YouTube. For example, in https://www.youtube.com/watch?v=RcmrbNRK-jY

RcmrbNRK-jY is the video ID.

Category: In general, you should first search for videos by category and then scrape them as it will save you a lot of time.

Please note that collecting at-least 1700 samples per category is mandatory.

Some tips:

1. Utilize packages like nltk to sanitize descriptions from Credits, contact information, subscription requests etc. as these things will not contribute towards better trained models or they may even exacerbate accuracy.
2. Try to minimize the effort you put into transforming b/w formats by only using just the above csv.
3. This exercise is for you to demonstrate your ability to handle large volumes of data. You are free to explore techniques online, but the code you submit should be completely written by you.
4. At the end of this exercise please submit the data through an excel sheet.

Text classification

You have to choose and use one model/techniques from *each* of the following categories

#Category	Model types
1	Linear classifiers, Naive-Bayes classifiers or SVMs
2	Bagging models, boosting models or shallow NNs
3	CNN, LSTM, GRU, Bidirectional RNNs, or RCNNS

Report

In the report you must submit

1. A Brief summary of how you scraped and pre-processed data from YouTube. Please note that candidates with original ideas and code written completely from scratch will be given more points.
2. The report *MUST* include a reason why you chose that particular technique or model for the classification.
3. For every technique you used, you *MUST* report precision, recall, and F1 Score. This should be in a graphical representation.
4. Finally, you will have to explain why you got the results and what has made that difference.

Final submission to be done at ashish@precily.com