CS109B – Advanced Topics in Data Science

Project Read me for Notebooks

Authors: Sathish Angappan, Hannah Bend, Yohann Smadja

Project Overview – Predicting Movie Genres

The overall theme of the final project is movie data with a focus on movie genre prediction. There were 5 milestones for this project and each milestone had a specific purpose (i.e. understanding the data, assembling training data, running machine learning algorithms etc.).

This readme gives details about the approaches tried in each milestone, how we progressed, what we found, where we stumbled and finally how we came here.

Milestone 1

In this Milestone, we focused towards understanding the APIs that were available to us (TMDB, IMDB, Wikipedia) and do some EDA (Exploratory Data Analysis)

We started off calling the different APIs and see what data is being returned and what can be used for our analysis. During Milestone1 we not only focused on the posters, but also tried to retrieve textual information for the movies (e.g.: Wikipedia overview, TMDB overview, movie cast / crew etc.).

We learnt that multiple posters are returned from the TMDB API and we decided to take the posters having at least 500px in width. Also, most posters were the same except for the color / different language texts.

We did some basic EDA like frequency of genres over the years, Budget Vs Revenue by Genre etc.

One major aspect we realized in this milestone was that almost all movies were classified into multiple genres. Since we always had 1-dimensional vectors for variables, we were not sure how to take care of the n-dimensions here. This turned out to be the multilabel classification which we started concentrating from Milestone 4.

Milestone 2

In Milestone 2, we concentrated mainly on different ways to vectorize the images. We analyzed on how to download movie posters, how to handle storage issues etc. While we downloaded posters, we had to introduce sleep() function to avoid getting locked by TMDB.

We had to decide on how to standardize the downloads across all movies, as many posters were returned for the same movie.

Also, we found that downloading posters of every movie released until now, is very time consuming and also could lead to storage issues. Also, too much data will take more time to process i.e. need higher GPUs. So, we decided to restrict our genres to only three (Comedy / Action / Drama) and downloaded movie posters for these genres only.

Milestone 3

For this milestone, we focused mainly on using traditional methods and analyze the predictions we received from these methods. We reduced the image dimensions using PCA and see how much variance is explained by PCA until 20-25 dimensions. For n_components = 20, around 65% of the variance was explained. We trained traditional models like KNN, Random Forests and logistic regression using these components and this did not give enough accuracy to make us happy.

We did basic keyword search and tried to get a pattern, so we could combine this pattern along with posters for our future milestones. Considering the amount of data, we had, this does not seem to be enough.

In Parallel, we started setting up AWS instances and Virtual Box and trying to figure out how to work with them. We faced multiple challenges and spent lot of time to make them work. Also, realized, the virtual box was using lot of memory even for running deep learning for small data sets. So we focused our attention towards AWS instances.

Milestone 4

This is when we had clear directions on how to proceed. We had identified what data we need to use. We had identified how to set up AWS instances, how to upload / download files, how to connect through putty. Moreover, we also tried creating Team AMI, so all of us can share the same instance for our analysis.

We did our first deep learning analysis using CNN / MLP and visualized the layers of CNN. We explored different approaches and had some good idea on how we will continue for Milestone 5.

We did some research on Multilabel classification. We did not have enough time to implement the same because time was spent on issues faced with AWS / other learnings.

Milestone 5

THE CLIMAX…….

With all the experience gained in the earlier milestones, we started directly by downloading 5000 movie posters, uploaded to AWS. We transformed the data to training / testing set and ran with pre-trained network (ResNet50) + SVM and created a simple model from scratch and analyzed the results. The Pre-trained network with SVM definitely gave much better results. The Pre-trained network seems to be a very optimized network. Since we did multilabel classification, this created imbalance in the dataset. As we added SVM with balanced weights, this was taken care and helped improve the accuracy.

And…. Stanley … was … happy ☺