Predicting movie ratings and viewer's sentiment analysis based on Movielens dataset

(By Manisha, Shwetha, Tanmay Sawaji, Vishwas Desai)

Background:

Manisha Chandran (manchand)

I, Manisha Chandran, have completed my undergraduate course of Bachelor of Engineering in Computer Engineering in 2019. In my undergrad course, I have taken up various courses such as Machine Learning, Operating System, Algorithms and Database Management System. Some of the projects I have worked on are:

- For my BE project, developed ML Algorithm for predictive maintenance of Heavy Commercial Vehicles (HCV), implemented web application for collecting data and manipulating it for analysis. Provided predictive analysis in graphical format and as a measure of percentage in JavaScript web application. This project was published in IJICS JOURNAL VOLUME 6. ISSUE 4. April 2019 (http://ijics.com/gallery/116-april-1076.pdf)

- Worked for NPCIL under the association of Bhabha Atomic Research Center during my summer internship on JavaScript based data validation library that was to be integrated with web applications developed within the organization.

- Worked for NPCIL during my summer internship to configure virtual private network and analyze web networks used within the organization. Used Datagram Transport Layer Security (DTLS) VPN protocol to solve the issues SSL/TLS has with tunneling over TCP.

Shwetha Panampilly (spanampi)

I, Shwetha Panampilly, have completed my undergraduate course of Bachelor of Engineering in Computer Science Engineering from SRM Institute of Science and Technology in 2019. After graduating, I have worked at INCAETEK as intern and Project Engineer with responsibilities including developing a pedestrian warning system, developing a mobile app, and developing user friendly interfaces for a CAD product, FreeCAD. Some of the projects I have worked in at INCAETEK are:

Identification of vendor from SMS with Payment details

- SMS messages containing payment details are parsed by an App developed using QPython and the geographical coordinates of the vendor are extracted. Google MAPS' Places API is used to identify the vendor.

Building extensions to support Part Development in FreeCAD

- INCAETEK has been engaged by Government of Kerala to conduct training programs for skill certification in modelling and prototyping functions of Product Development. FreeCAD tool is planned to be used in this initiative. While FREECAD is functionally rich, there is a need to extend its usability by developing friendly user interfaces. This project is to build extensions to FreeCAD to enhance user friendliness at par with industry standard CAD software packages. Development is done using Python.

Pedestrian warning system for no-pedestrian zones

- A computer with GPU accelerator connected to cloud via Wi-Fi/ethernet and a camera are deployed. The images from the camera are processed real time with cloud computing support. The actuation element (buzzer, speaker, warning lights etc.) are connected to a microcontroller connected to internet through Wi-Fi/ethernet.

Tanmay Sawaji (tsawaji)

I, Tanmay Sawaji, have completed my undergraduate course of Bachelor of Engineering in Computer Engineering in 2017. After graduating, I have worked at Aniruddha Telemetry Systems (ATS) as a Machine Learning Engineer with responsibilities including computer vision, machine learning, server maintenance, server scripting, creating APIs, automation, and single-board computer programming. Some of the projects I have worked on, at ATS, are:

- Automated Number Plate Recognition (ANPR) for coal trucks where visibility, clarity and language of the number plate may differ.
- Outlier Capsule Detection Machine (OCDM) where the task is to filter out outlier capsules from a batch based on color, size, shape, and defects.
- Controller for Driver Authentication system at an Aluminum plant which manages API calls to face detection, breathe alcohol analysis and palm vein reader to authorize a driver.
- In-house attendance system that marks attendance based on a facial recognition API and maintains and generates reports.
- Setup of a NTP server that sets network time based on GPS data instead of a global NTP server pool.

Vishwas Desai (visdesai)

I, Vishwas Desai completed my Bachelor of Engineering degree in Information Science Engineering in 2020. My coursework included subjects like Python, Artificial Intelligence and Machine Learning. In my final year of undergrad, I did a project on predicting Traffic Congestion using Machine Learning which involved integration of embedded systems and ML techniques. Apart from my undergrad coursework, I have completed a few MOOC courses involving Python and ML namely Python for Everybody specialization and Machine Learning with Python offered by IBM. I have also enrolled myself in Machine Learning MOOC course offered by Stanford University and taught by Andrew Ng which will be completed by June 2021.

Project description:

Our project uses the movielens dataset. This dataset contains key attributes like movie titles, genres, ratings, IMDb ID, TMDb ID. Our primary goal is to predict ratings of a movie based on the above attributes and additional attributes will be called by making use of the IMDb API. The secondary goal is to binarize the reviews attribute by applying sentiment analysis which is trained on different dataset containing reviews and sentiment. Through this project we intend to predict the performance of the movie and the viewer's sentiment.

Motivation

The main aim is to assist the users to make a calculated choice of whether to stream the movie or not. This also can be used by various streaming platforms like Voot, Hotstar etc. to gauge the movies they want on their platform.

Technical description:

The plan is to perform regression using an artificial neural network (ANN) with ratings attribute as our target variable. Attributes like cast, director, reviews, budget, box office earnings will be obtained using the IMDb API. We plan to find a correlation between the above attributes with our target variable to determine the viability of our attributes. We plan to convert the reviews

attribute to a continuous value by determining the sentiment of all reviews and calculating the percentage of positive reviews for a specific movie. We will train another probabilistic model to determine review sentiment which uses a different dataset i.e., IMDb 50K dataset.

Data Description

The datasets being used include Movielens Dataset, IMDb 50K dataset are available on grouplens.org and Kaggle respectively and an additional dataset to retrieve the additional attributes by using the IMDb API.

References

[1] MovieLens

[2] IMDb 50K

[3] https://www.sciencedirect.com/science/article/pii/S1110866516300470

[4] https://builtin.com/data-science/recommender-systems

[5] IMDb API

Potential challenges/hurdles:

- Extensive preprocessing of the raw dataset required.
- Uncertain on how to implement an ANN for Regression.
- A part of the group is new to Machine Learning.
- Hardware limitations
- IMDB API has a limit of 1000 API calls per day and our dataset has around 10,000 movies.