Analysis on IMDb Database

Group 4

December 16, 2017

Contents

1 Group Information

English Name	Chinese Name	Student ID
Jeff	傅永升	1430003004
Covey	刘克盾	1430003011
Garfield	邬嘉祺	1430003029
Frank	邬可夫	1430003030
Bill	钟钧儒	1430003045

2 Project Introduction

In this section, information of the project will be introduced.

2.1 Project Title

Analysis on IMDb Database.

2.2 Goals

By analyzing the dataset from the IMDb, find some associations between columns, classcify movies according to some criterias.

3 Data & Data Collections

In this section, information of the dataset used in this project will be introduced.

3.1 Dataset

Here are some information of the dataset itself.

- Formate: csv file.
- Size: 302KB
- Number of rows: 1000
- Number of columns: 12
  - Rank
  - Title
  - Genre
  - Description
  - Director
  - Actors
  - Year
  - Runtime
  - Rating
  - Votes
  - Revenue
  - Metascore
- Source: IMDb (https://www.imdb.com)

3.2 Data Collection

Internet Spider The most common way to collect such dataset is to use Internet spiders. In this case, a Python packge called Scarpy was used.

4 Analysis

Raised by members in the group, attributes are analyzed by members individually.

4.1 Analysis on Directors

The section is provided by Junru (Bill) Zhong. This analysis aims to find out the following associations.

- Find out which director earned the largest revenue.
- Find out the gerene of films of the first three directors with the largest revenue.
- Estimate how much they will earn for their next film.

4.1.1 Code implementation

In this section, Python (Anaconda) was used, with package Pandas.
The steps are listed below.

1. List all directors.
2. Get the total revenues each directors got.
3. Rank all directors by revenues got each film in average.
4. List information of directed film(s) of the top-3 directors.
5. Estimate the revenues of their next film by some algorithms.

4.2 Revenue and Genre

This section is provided by Yongsheng (Jeff) Fu. This analysis is to find out the change in taste of genre from 2006 to 2016.