Analysis on IMDb Database

Group 4

December 13, 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.