Group Project:
Analysis on IMDB database

Introduction

There have been several interesting studies on *movie recommendation* in the past years. For example, many different data mining methods were combined and performed to find the best suggesting model for Netflix Challenge. To achieve an acceptable recommending system, movies have to be categorized based on their similarities. Even for human experts this task is not trivial, and requires a comprehensive analysis.

There are several key features to identify similar movies, such as their genres, IMDB scores, or by analyzing the cast of the movies. In this project, you are requested to deliver a *genre prediction*, *interesting rules detection* for casts (which groups are normally seen together), and segmentation of *similar* movies.

Steps in the Project

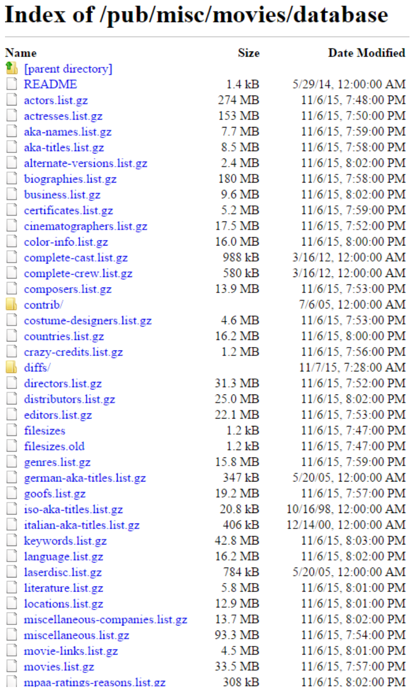
You are responsible for every step of this analysis, which are:

- Dataset creation.
- Genre prediction (classification)
- *The usual* casts (association rule mining)
- Finding similar movies (clustering)
- Report

The details are explained in the following sections.

Dataset creation

IMDB database contains an overwhelming amount of information about movies. However, to accomplish the requested objectives, using *all* of such data is impractical.



Index of /pub/misc/movies/database

Name	Size	Date Modified
[parent directory]		
README	1.4 kB	5/29/14, 12:00:00 AM
actors.list.gz	274 MB	11/6/15, 7:48:00 PM
actresses.list.gz	153 MB	11/6/15, 7:50:00 PM
aka-names.list.gz	7.7 MB	11/6/15, 7:59:00 PM
aka-titles.list.gz	8.5 MB	11/6/15, 7:58:00 PM
alternate-versions.list.gz	2.4 MB	11/6/15, 8:02:00 PM
biographies.list.gz	180 MB	11/6/15, 7:58:00 PM
business.list.gz	9.6 MB	11/6/15, 8:02:00 PM
certificates.list.gz	5.2 MB	11/6/15, 7:59:00 PM
cinematographers.list.gz	17.5 MB	11/6/15, 7:52:00 PM
color-info.list.gz	16.0 MB	11/6/15, 8:00:00 PM
complete-cast.list.gz	988 kB	3/16/12, 12:00:00 AM
complete-crew.list.gz	580 kB	3/16/12, 12:00:00 AM
composers.list.gz	13.9 MB	11/6/15, 7:53:00 PM
contrib/		7/6/05, 12:00:00 AM
costume-designers.list.gz	4.6 MB	11/6/15, 7:53:00 PM
countries.list.gz	16.2 MB	11/6/15, 8:00:00 PM
crazy-credits.list.gz	1.2 MB	11/6/15, 7:56:00 PM
diffs/		11/7/15, 7:28:00 AM
directors.list.gz	31.3 MB	11/6/15, 7:52:00 PM
distributors.list.gz	25.0 MB	11/6/15, 8:02:00 PM
editors.list.gz	22.1 MB	11/6/15, 7:53:00 PM
filesizes	1.2 kB	11/6/15, 7:47:00 PM
filesizes.old	1.2 kB	11/6/15, 7:47:00 PM
genres.list.gz	15.8 MB	11/6/15, 7:59:00 PM
german-aka-titles.list.gz	347 kB	5/20/05, 12:00:00 AM
goofs.list.gz	19.2 MB	11/6/15, 7:57:00 PM
iso-aka-titles.list.gz	20.8 kB	10/16/98, 12:00:00 AM
italian-aka-titles.list.gz	406 kB	12/14/00, 12:00:00 AM
keywords.list.gz	42.8 MB	11/6/15, 8:03:00 PM
language.list.gz	16.2 MB	11/6/15, 8:02:00 PM
laserdisc.list.gz	784 kB	5/20/05, 12:00:00 AM
literature.list.gz	5.8 MB	11/6/15, 8:01:00 PM
locations.list.gz	12.9 MB	11/6/15, 8:01:00 PM
miscellaneous-companies.list.gz	13.7 MB	11/6/15, 8:02:00 PM
miscellaneous.list.gz	93.3 MB	11/6/15, 7:54:00 PM
movie-links.list.gz	4.5 MB	11/6/15, 8:01:00 PM
movies.list.gz	33.5 MB	11/6/15, 7:57:00 PM
mpaa-ratings-reasons.list.gz	308 kB	11/6/15, 8:02:00 PM

For the first step, you should carefully read and understand what you are asked for. Once you understand the problem, you know what features you should include in your dataset to tackle the challenges. You should clearly state what features could be beneficial in your analysis and how you manage to obtain them. Remember, you *cannot* achieve good results if you do not have a good dataset.

There is no specification for your dataset, and you can create whatever feature you desire. In fact, creating informative features are essential in this part. However, your dataset should have:

1. At least 10000 random movies.

2. *Title*, *Year*, and *Genre* as features. You may (and should) add several other features, but these features are necessary for identification and evaluation.

For this step, you may use any programming language and technique you know to construct the dataset in .csv format; it is not necessary to use R. For your demo, you should provide this file. It is *very* unlikely that two different groups create the same dataset or features.

The main portion of this part's marks is dedicated to the features you construct. Do some research, think carefully, and come up with interesting ones!

Genre prediction

After creating your dataset, set aside 20% of your movies for test, and use the other 80% for training. Use the first digits of your UFIDs for seed for result reproduction purposes. For later convenience, save them in separate .csv files. You *might* be asked later to run your predictive model on the test set.

Genre is used as the target variable for the classification part. Based on your knowledge of different classification methods and the dataset, you should be able to justify the results and answer the following questions:

- What were your options for classification?
- How did you evaluate different classification techniques?
- What measures have you taken to improve the results?

The usual casts

Who would normally work together? We ask.

Find the first 20 nonredundant groups (*rules*) sorted based on their support and confidence. The groups should contain at least 4 people; director, author, producer, actor, actress, and more. The answer to this part depends on

- The number of different roles you have in your dataset.
- What you put in the left-hand-side of the rules.
- How you would automatically check all the possibilities.

Moreover, provide some rules which indicate the relation between *genre* and the *cast*. You should explain why this finding is aligned (or not) with your results for the *genre prediction* part.

Finding similar movies

For this part, you should cluster movies into k clusters. Here k is the number of different genres you have in your dataset. If your dataset contains both nominal and numeric attributes, it is your responsibility to develop a clustering method which works for this case.

You will be asked:

- Which clustering method works best in this case? And why.
- Would clustering the movies into k' clusters where $k' > k$, help in better categorization? And how.

Room for improvement?

After taking all the steps, the deficiencies become obvious. By now, you should know that for a given dataset, you cannot improve your methods further than a certain amount; no matter how much you tune your method parameters.

Go back to the first step, and add or adjust whatever you need to improve your results. Document every step you take to show your starting point, challenges and solutions.

Report

The last but not least, you should provide a report in an article format. In the end, each group should present their report which covers the following:

- Literature review: read and summarize what others have done on this data or subject. Clearly state the differences and similarities to your work and the ideas that you leveraged from their approaches.
- Brief statistics about your dataset: Provide an informative summary about the dataset(s) you created. What it covers, how many different genres, directors,... it has and etc.
- Method and Materials: Every step of the way you take for your project should be included in your report. This part should cover a qualitative description about different parts of the project.
- Results: Bring your results from different parts in your report.
- Discussion: Justify your results; why you think your results are impressive!
- Conclusion: Conclude your report in one or two paragraphs.
- References: Articles or works that you review or use for this project.

Grading

For grading, the report is explained by group members. If there is any ambiguity in any part of the report, you will be asked to show and run your code as a validity check for the methods and results.

There is no specific threshold for each part's accuracy (or performance). The effort for understanding the problem, providing the best solutions, and making improvements are the most important aspects of this project.