Department of Computer Science

bu.edu/cs

Incluvie : Actor Data Collection

Ada Gok, Dana Hochman, Lucy Zhan

Introduction and Task

Incluvie is a platform that promotes and celebrates diversity within Hollywood. The platform allows movie viewers to voice their opinions on a movie regarding diversity.

Our task is to create a method that extracts actor data, including gender and cultural background. Incluvie will use our method and the generated datasets to display the level of gender and cultural diversity of movie casts. In addition, we will use the dataset to analyze any possible correlation between the gender and race of an actor and his/her popularity.

Figure 1. Incluvie Logo

Approach

Our main method of data extraction takes a text file of actors' names listed line-by-line as input.

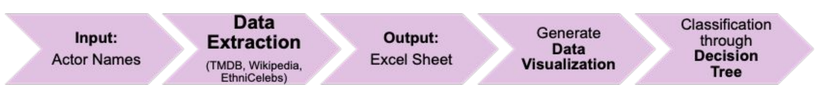
For each actor's name, we use the TMDB API, Wikipedia API, Ethnicelebs and BeautifulSoup to get data. **TMDB** extracts the gender, birth place and popularity score of the actor. **Wikipedia** extracts an ethnicity list from the actor Wikipedia page using keywords: 'ancestral', 'descent', 'ancestry', 'mother', 'father', 'mom', 'dad', and 'parents'. **Ethnicelebs'** pages have a specific section that lists out the ethnicities of a given actor, so we retrieve it and clean up the string to only consist of the ethnicity titles.

For Wikipedia and Ethnicelebs, we reference a text document of common ethnicities that is converted into a list of strings in Python and used it to detect ethnicity keywords on the webpage. For scraping data off the webpages, we mainly use the **BeautifulSoup** package that parses the HTML code.

The output is a data frame containing all the data parameters, which is then exported as an excel sheet.

We then proceed to generate data visualization to showcase the gender and cultural distribution of Hollywood using the Python library **Matplotlib**.

We choose to implement decision tree to classify an actor's popularity as low, medium, or high based on the feature dimensions: actor's race and gender. We split the popularity range into thirds and that will be the range of the three popularity statuses. The input data for the decision tree consists of binary variables {0, 1} to represent whether an actor is of a certain race or gender. The decision tree trains 80% of the data and tests on 20%.



Dataset and Metric

We run data extraction on two sets of actors: Top 100 Highest Grossing Stars of 2018 [1] and Top 100 IMDB Stars via Starmeter [2].

The data parameters we focus on are gender, birth place, ethnicities from Wikipedia, ethnicities from Ethnicelebs, TMDB popularity score (1 as the lowest score), and percent similarity/difference as a comparison parameter for the two ethnicity lists. Two more parameters we manually add are race (sourced -- meaning race determined by ethnicity lists) and race (gut check -- meaning race determined by guess*). The gut check is for data accuracy check.

*We realize there may be bias involved.

Name	Gender	Birth Place	Ethn. (Wiki)	Ethn. (EC)	Pop. Score	% similar	% diff.	Race Source	Race Gut Check
Dwayne Johnson	male	Hayward, CA, USA	Irish, Nova Scotian, Samoan	Nova Scotian, Samoan, Irish, African Canadian	11.538	75	25	African American, White, Pacific Islander	African American, White
...	...	...	...	...	...	...	...	...	...

Figure 2. Example of Dataset

Results

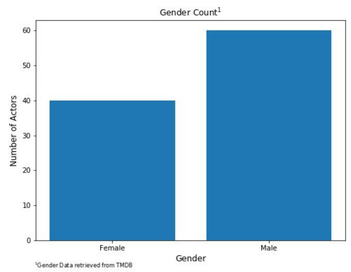


Figure 3. Gender Bar Graph for Top 100 Highest Grossing Stars 2018

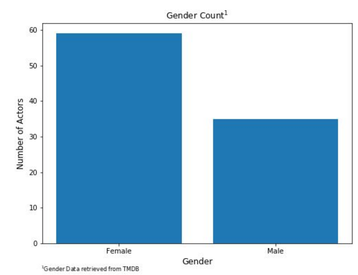


Figure 4. Gender Bar Graph for Top 100 IMDB Stars via Starmeter

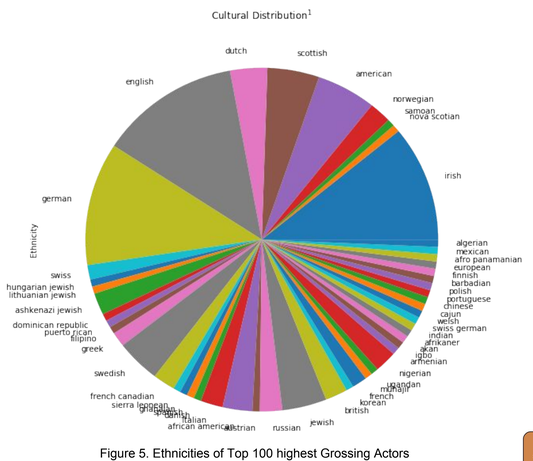


Figure 5. Ethnicities of Top 100 highest Grossing Actors

We generate multiple graphs that showcase the cultural distribution among the actors. Figure 5 is an example of a cultural distribution pie chart of the Top 100 Highest Grossing Stars of 2018.

We implement the decision tree for all 200 data points of the combined Top 100 Highest Grossing Stars and Top 100 IMDB Stars. For our decision tree test, we receive a high accuracy of around 90%.

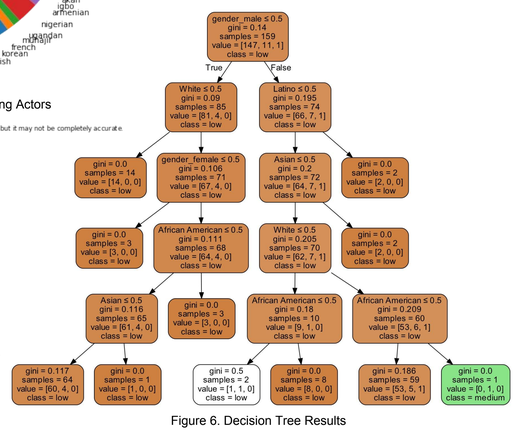


Figure 6. Decision Tree Results

Conclusion

From the results, we learn a lot about gender and ethnic diversity in Hollywood. From Figure 3 and Figure 4, we conclude that the highest-grossing stars (according to Top Highest Grossing Stars [1]) have a 60:40 male to female ratio, but the rising stars (according to the IMDB Starmeter [2]) have a lower male to female ratio. Thus, female actors are rising in popularity, but male actors are paid more.

From Figure 5, we identify that the majority of actors in Hollywood are of European or American origins.

Although the decision tree has 90% accuracy, we want to disclose that the dataset is too small to give a truly accurate prediction of popularity status (low, medium, or high). In addition, we can't completely rely on the TMDB popularity score considering that 91% of the top Hollywood actors have scores in the low range.

Future Work

The current data extraction method is not optimized in terms of time efficiency. It will be time consuming to run the dataset in the future for a larger set of actors. Refactoring the code to run faster is necessary.

Additionally, we hope to make the data extraction dynamic. Currently the code takes in a text file as input. In the future, we want to find a reliable list of top 100 actors that is automatically updated daily/weekly. As the list updates, our aim is for the dataset to be recompiled and updated to showcase data visualization for the most recent top 100 actors.

References

1. "Highest Grossing Stars of 2018 at the Domestic Box Office." The Numbers - Where Data and Movies Meet. Accessed December 03, 2018. https://www.the-numbers.com/box-office-star-records/domestic/yearly-acting/highest-grossing-2018-stars.
2. "IMDb Top Star Meter." IMDb. Accessed December 03, 2018. https://m.imdb.com/chart/starmeter.