Pavitra Dangati

Manan Patel

Thomas Lam

COMPSCI 590V

Selection of Data and Problem Assignment

**Topic**: Analysis of Diversity in Hollywood Movies

**Description:**

Ethnic and gender diversity has been a long standing issue that is gaining light over the recent years in various sectors, especially in Film, Arts and Theatre. Although these diverse groups are mentioned in a lot of the work over the years, the minority certainly feels that they are misrepresented. For this project, we want to look at a subset of the creative outlet, that is, Movies by Hollywood and perform an analysis of racial and gender makeup of the cast and crew of 5000 movies from 1960-2016 timeframe.

The dataset we’re looking at is from Kaggle [1], which contains information on different movie attributes such as directors, durations, actors, ratings, genres, etc. We’re planning to supplement this dataset with racial and gender makeup data on casts and crews that appear in the Kaggle dataset. We plan to scrape Notable Names Database and leverage Python module SexMachine to find these additional racial and gender attributes. We’re also using historical datasets from Wikipedia, Data.gov and Worldbank ([3], [4], [5]) websites on racial and gender makeup of US population in our analysis.

Using these combined datasets, we will perform a comparative analysis on whether gender and racial ratios in Hollywood movies accurately reflect those in general US population. Our hypothesis is that there is a statistically significant difference between those ratios which is indicative of the diversity problem in Hollywood. We are also interested in exploring how these ratios change over time. We hope to find the best visual representation that will allow us to easily summarize this analysis in an intuitive manner.

**References:**

[1] Kaggle dataset: <https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset>

[2] Notable Names Database: <http://www.nndb.com>

[3] Wikipedia article on racial and ethnic demographics of the US: <https://en.wikipedia.org/wiki/Historical_racial_and_ethnic_demographics_of_the_United_States>

[4] Data.gov website for datasets on breakdown of US by race and gender:

<https://catalog.data.gov/dataset?q=population>

[5] Worldbank data

<https://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS?locations=US>

Pavitra Dangati

Manan Patel

Thomas Lam

COMPSCI 590V

Initial Project Description

**Topic**: Analysis of Diversity in Hollywood Movies

**Description:**

Ethnic and gender diversity has been a long standing issue which is gaining light over recent years in various sectors, especially in Film, Arts and Theatre. Although US demographic has become more diverse in recent decades, there’s been various studies which show that minority and female representation in the arts is severely lacking given such demographic trend. For this project, we were looking something for fun visualization on many sources and we found a data set on diversity in Hollywood and we also wanted to look at a subset of the creative outlet, that is, Hollywood movies and perform an analysis of racial and gender makeup of the cast and crew of movies from 1960-2016 timeframe. Now if we consider top 200 movies released in 2016 and cable and digital platform tv shows from 2016-17 in order to which women and people of color are present in front and behind the camera.

We want to provide a visual aid that helps clearly illustrate this problem to a targeted audience consisting of Hollywood executives, decision makers, directors, scriptwriters as well as other people interested in exploring this social issue. Here are some questions that hopefully our visualization design can help answer:

1. How many female and minority actors/actresses are typically represented in movies compared to their white, male counterparts?

2. How did this representation change over time?

3. What about behind the camera representation? i.e. directors, writers, crews, etc.

4. Is this diversity problem unique to only a select few genres or is this a universal phenomenon across all genres?

5. Do more popular and mainstream movies (with various popular metrics such as facebook likes, IMDB scores, total revenue, etc.) tend to be less diverse than less popular ones?

6. The popular directors who have a bigger platform and opportunities to present the under represented actually have high minority representation in their movies?

7. Did the movies with high minority representation do well in the international markets?

The dataset we’re looking at is from Kaggle [1], which contains information on different movie attributes such as directors, durations, actors, ratings, genres, etc which is in csv format. We’re planning to supplement this dataset with racial and gender makeup data on casts and crews that appear in the Kaggle dataset. We are scraping the Notable Names Database and leveraging Python modules to find these additional racial and gender attributes. These fields would be added to our master movie dataset from kaggle. We’re also using historical datasets from Wikipedia, Data.gov and Worldbank ([3], [4], [5]) websites on racial and gender makeup of US population in our analysis. All these data would also be loaded in csv formats.

Using these combined datasets, we will perform a comparative analysis on whether gender and racial ratios in Hollywood movies accurately reflect those in general US population. Our hypothesis is that there is a statistically significant difference between those ratios which is indicative of the diversity problem in Hollywood. We are also interested in exploring how these ratios change over time. We hope to find the best visual representation that will allow us to easily summarize this analysis in an intuitive manner.

To perform analysis and to gain insights to answer the questions that the users might ask, we need to merge various data sources, group, order and filter values, derive new and cumulative values for comparison, determine what visualisations would help in answering the questions better, form general hypothesis for the data and prove it correct or wrong by further exploration.

Although this is a well-studied phenomenon, we think that visual aid helps highlight this issue with more impact. We hope that our visual design will be used to bring about positive casting decisions that are more suited for the times, bring about stories from different cultures to a wider audience and provide employment opportunities to the minorities.

As we are performing a study over the diversity trends in the previous years and their real world representation, we are using visualisation techniques over machine learning and statistical analysis as we aren’t using the current data to determine information like how likely is the minority representation to increase in the next few years and by how much. Visualisation aids in analysis of data, to get rich patterns or finding anomalies from the given dataset unlike machine learning where we try to predict new information accurately.

**References:**

[1] Kaggle dataset: <https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset>

[2] Notable Names Database: <http://www.nndb.com>

[3] Wikipedia article on racial and ethnic demographics of the US: <https://en.wikipedia.org/wiki/Historical_racial_and_ethnic_demographics_of_the_United_States>

[4] Data.gov website for datasets on breakdown of US by race and gender:

<https://catalog.data.gov/dataset?q=population>

[5] Worldbank data

<https://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS?locations=US>