The Character Impact Project

Leveraging Machine Learning Techniques to Measure Character Screen Time on Episodic Television Content

Alden Chico

720,800,000,000$

The Projected Net Worth of the Entertainment Industry by 2020 (Statista)

Problem

How can entertainment companies maximize the value of their episodic show content?

Goal

Craft compelling storylines that lead to a show’s seasonal renewal.

Idea

Measure character screen time using machine learning techniques to make better decisions for the television show.

Project Proposal

Use machine learning to measure how much time each character from *The Office* Season 1 spends on screen. Develop a mock IMDb page that showcases this information for each episode of the show.

Gathering Audience Feedback from the IMDb Dataset

Objective: Filter the IMDb Dataset for information related to *The Office* and the user ratings associated with Season 1 of the show.

IMDb Dataset

title.basics
tconst
titleType
primaryTitle
originalTitle
isAdult
startYear
endYear
runtimeMinutes
genres

title.episode
tconst
parentTconst
seasonNumber
episodeNumber

title.ratings
tconst
averageRating
numVotes

title.principals
tconst
ordering
nconst
category
job
characters

name.basics
nconst
primaryName
birthYear
deathYear
primaryProfession
knownForTitles

Data Wrangling

title.basics

tconst
titleType
primaryTitle
originalTitle
isAdult
startYear
endYear
runtimeMinutes
genres

title.basics: Basic information for all titles contained in the Internet Movie Database

- tconst (string): Alphanumeric unique identifier for title
- primaryTitle (string): The title associated with the media

Step 1: Find the unique identifier (*tconst*) associated with *The Office*.

Data Wrangling

title.episode
tconst
parentTconst
seasonNumber
episodeNumber

title.episode: TV Episode information

- tconst (string): Alphanumeric identifier for the episode
- parentTconst (string): Alphanumeric identifier of the parent TV series
- seasonNumber (integer): Season the episode belongs to
- episodeNumber (integer): Episode number of the tconst in the TV series

Step 2: Find *tconst* for each episode of *The Office* Season 1 and use the *title.basics* dataset to find the name of each episode.

Data Wrangling

title.ratings
tconst
averageRating
numVotes

title.ratings: IMDb rating and votes information for each title in the dataset

- tconst (string): Alphanumeric identifier for the title
- averageRating: Weighted average of all the individual user ratings

Step 3: Gather the episode’s user rating by referencing *tconst* for each episode.

Data Wrangling

title.principals
tconst
ordering
nconst
category
job
characters

title.principals: Principal cast and crew for titles

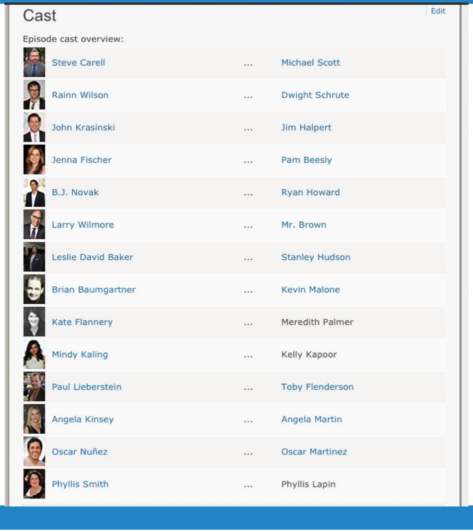
- tconst (string): Alphanumeric identifier for the title
- nconst (string): Alphanumeric identifier for name/person
- category (string): Category of job person was in
- characters (string): Name of the character played

Step 4: Find the principal character information from each episode by referencing each episode’s *tconst*.

Problem

title.principals only contains information for **principal** characters of the show. We want to retrieve information about **every** character in the show.

Data Wrangling



Cast

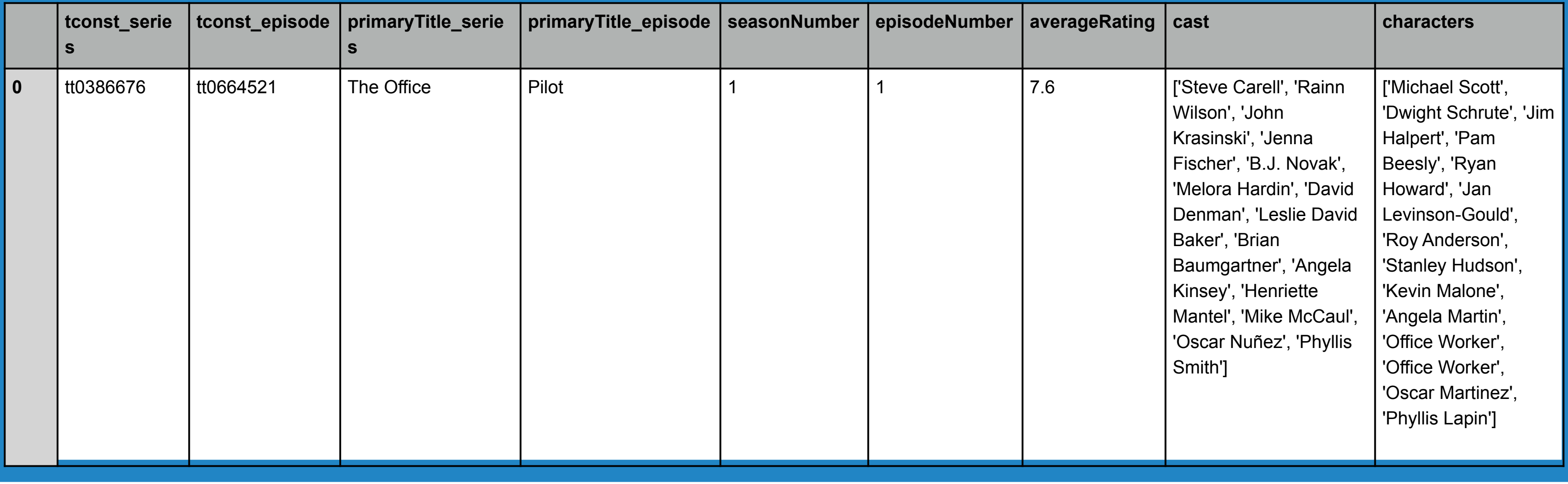
Edit

Episode cast overview:

Steve Carell	...	Michael Scott
Rainn Wilson	...	Dwight Schrute
John Krasinski	...	Jim Halpert
Jenna Fischer	...	Pam Beesly
B.J. Novak	...	Ryan Howard
Larry Wilmore	...	Mr. Brown
Leslie David Baker	...	Stanley Hudson
Brian Baumgartner	...	Kevin Malone
Kate Flannery	...	Meredith Palmer
Mindy Kaling	...	Kelly Kapoor
Paul Lieberstein	...	Toby Flenderson
Angela Kinsey	...	Angela Martin
Oscar Nuñez	...	Oscar Martinez
Phyllis Smith	...	Phyllis Lapin

Step 5: Use beautifulsoup to parse through the casting tables of each episode's webpage from Season 1 of *The Office* to retrieve cast/character information for the show.

Data Wrangling



	tconst_series	tconst_episode	primaryTitle_series	primaryTitle_episode	seasonNumber	episodeNumber	averageRating	cast	characters
0	tt0386676	tt0664521	The Office	Pilot	1	1	7.6	['Steve Carell', 'Rainn Wilson', 'John Krasinski', 'Jenna Fischer', 'B.J. Novak', 'Melora Hardin', 'David Denman', 'Leslie David Baker', 'Brian Baumgartner', 'Angela Kinsey', 'Henriette Mantel', 'Mike McCaul', 'Oscar Nuñez', 'Phyllis Smith']	['Michael Scott', 'Dwight Schrute', 'Jim Halpert', 'Pam Beesly', 'Ryan Howard', 'Jan Levinson-Gould', 'Roy Anderson', 'Stanley Hudson', 'Kevin Malone', 'Angela Martin', 'Office Worker', 'Office Worker', 'Oscar Martinez', 'Phyllis Lapin']

Example Result from Data Wrangling the IMDb Dataset

Interpreting the IMDb Dataset Information

Objective: Create visualizations to help us understand the data we gathered from the IMDb dataset.

Available Information

♀ ♂

Gender

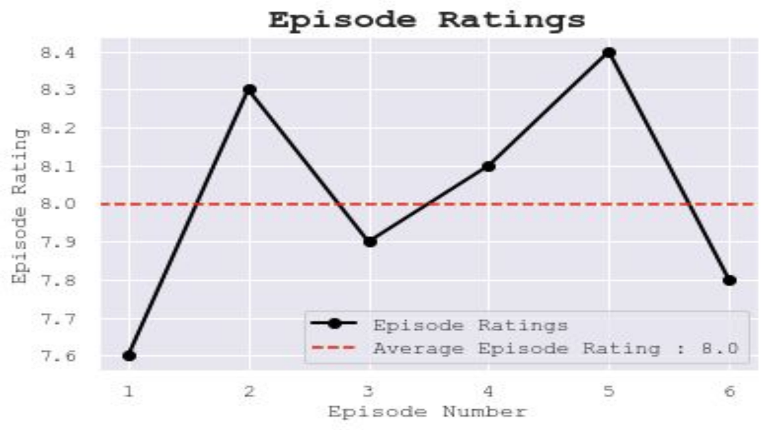
Age

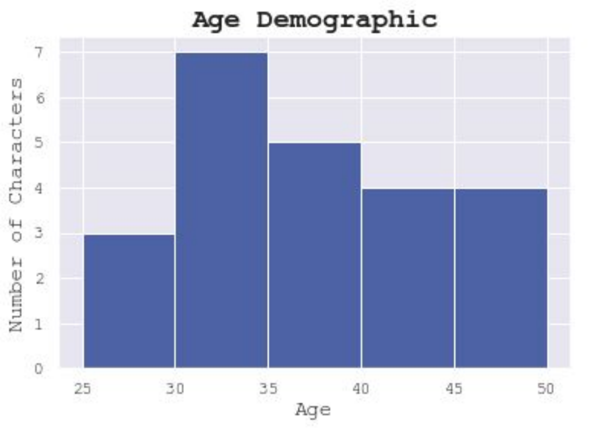
Ethnicity

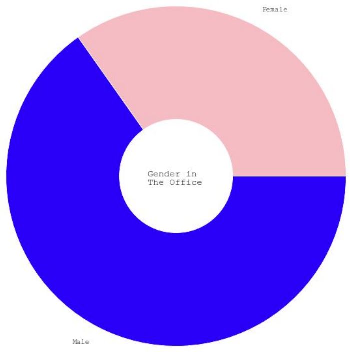
☆

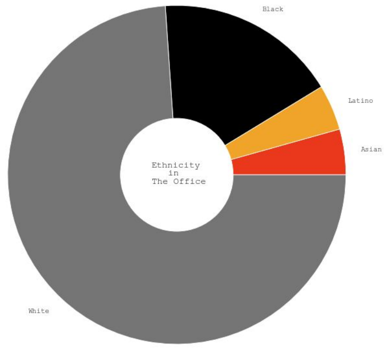
Episode Rating
Number of Character Appearances

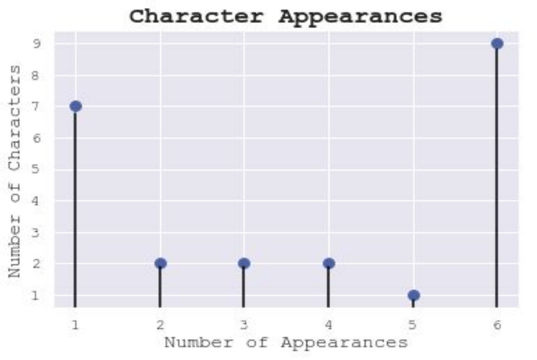
Exploratory Data Analysis











Takeaways

Episodes 1, 3, and 6 were rated below average among Season 1 episodes of *The Office*.

9 characters showed up in the casting table for every episode from Season 1.

Takeaways

The character age is fairly spread out between 25 and 50 years of age.

White men make up more than 2/3rds of the cast in the show

Main Takeaway

Cast homogeneity may affect the performance of our model's facial recognition classifier. We should be aware of this moving forward.