CS 3300 Project 2: Movie Dialogue

Elaine Hwang (yh467), Katerina Prastakou (kmp244), Brianna Singer (bls252)

DATA DESCRIPTION

Our project showcases commonly used dialogue in films, a.k.a. cliches, and their relations to average rating and genres of the movies they are used in. Due to the fact that there was no database provided for our specific project, we created the relational database to be used.

The first dataset is a set of strings which represent the cliches and unique IDs.

The second dataset is a collection of all movies in which the cliches appear. This information was collected from scraping QuoDB, an online database of movie screenplays with information about thousands of movies based on on-screen dialogue.

From this database we collected information on each common line of dialogue as well as the titles and quantities of films which include that quotation or a variation thereof. We noted which quotations had comparatively small representation in films and removed them from the dataset. Ou final cliche selection includes 23 popular quotes.

This database also included the genres and overall critical rating of each film, information that was scraped from iMBD.

Overall, our tables were as follows:

Cliches	Movies	QuoteMovie
clicheID (unique int)	movieID (unique int)	clicheID (int)
cliche (string)	movieTitle (string)	movieID (int)
	genres (string)	
	rating (float)	

We included columns to describe information about each line of dialogue such as the total count of films which include it, the average iMDB rating of each set of films, and the genres of each film in the set.

VISUAL ELEMENTS

We square root scaled the total count of films for each line of dialogue to the size of the font inside the bubbles. We linear scaled rectangles in the histogram based on the count of film rating for each line of dialogue. We did not scale the number of bubbles in each genre cluster, instead we based them on the number of films that fall into that genre per line of dialogue.

We made a slider to filter out the dialogue bubbles by average rating so that the user can easily see that most cliched dialogue is used in films with middle ratings.

We made the dialogue bubbles change color for emphasis when clicked so that is is clear for which one the user is seeing data.

We added hover bubbles in the genre chart so that the user knows which movie the bubble is referring to.

THE STORY

You can see from our visualization that cliched dialogue is used across genres, but especially in comedy and drama films. Also, films that use cliched dialogue tend to have ratings which peak around 5 or 6 out of 10. Perhaps this means that audiences enjoy repeatable dialogue that includes a familiar joke or an easily digestible character motivation, but movies which rely on these do not win especially high reviews in the long-run.

CITATIONS

Brianna:
A Stack Overflow Question About How to Add Text to Circles
A YouTube Tutorial with an Introduction to Forces in d3
A Stack Overflow Question About Changing Styling Programmatically in D3

Elaine:
A Brief Tutorial on onClick events
A Stack Overflow Question About Animating SVG Rectangles
iMBD: A Database of Movie Information

Katerina:
A Stack Overflow Question About Repositioning Force Nodes in d3
A Coded Example of an Animated Bar Chart in d3
A Database of Movie Screenplays Called QuoDB
A Plugin For Showing Data in d3 Upon Mouseover