The Simpsons

CS 3300 Project 2

By Vaidehi Garg (vg254), Brendan Fox (bf264), Jason Kwon (yk453)

Story

What does your visualization tell us? What was surprising about it?

The Simpsons has been on the air since December 17, 1989, with “Simpsons Roasting on an Open Fire.” Its 29th season began on October 1, 2017, and is still being broadcast. The Simpsons had been one of the most popular TV programs in this country with more than 500 episodes.

The idea behind our visualization was motivated by the fact that the “popularity”, which we measured with the number of viewers of a certain episode, does not necessarily indicate the “quality” of the episode, which could be measured by its IMDb rating. Another problem that seems to arrive often is that people have trouble deciding which Simpsons episode to watch, since there are SO MANY of them. We intend for our webpage to be helpful to those who are looking for interesting Simpsons episodes to watch, and make their decision-making process just a little easier.

In order to figure out which factors were contributing to those two variables, we decided to analyze each episode in terms of how often certain characters spoke, and how often those lines were spoken in certain locations. These two factors are displayed as the user selects a season with the slidebar and an episode with the left and right buttons. This enables the user to easily access the episode they watched and liked the most in the past and check which characters were taking the primary position in terms of the number of spoken lines in that particular episode and where they were spoken. Our data visualizations can also be used by a user to find an episode they think they should watch next!

More details in the following sections.

Data

A description of the data. Report where you got the data. Describe the variables. If you had to reformat the data or filter it in any way, provide enough details that someone could repeat your results. If you combined multiple datasets, specify how you integrated them. Mention any additional data that you used, such as shape files for maps. Editing is important! You are not required to use every part of the dataset. Selectively choosing a subset can improve usability. Describe any criteria you used for data selection.

For our visualizations, we used two datasets, both compiled by Todd Schneider for his post on The Simpsons. The datasets were created by scraping several websites, the primary of which was Wikipedia. The datasets are available for public use on Kaggle.

The first dataset we used is called simpsons_episodes.csv. It contains various details about each of the episodes of The Simpsons, as shown in the table below.

Header	**Definition**
season	The season the episode is part of
number_in_season	The episode’s number in the corresponding season
us_viewers_in_millions	Total number of viewers in the US, in millions
imdb_rating	IMDb rating of the episode

This dataset was used to plot the first visualization, which simultaneously shows the trends of the viewership and the IMDb rating of each episode in a given season. The season can be selected using the donut slider included below the graph.

This dataset was also used in the “second part”, which displays the title and a screengrab of each episode of the selected season, along with a link to watch the episode. Users can scroll through the episodes in a season using the buttons provided. The variables from the dataset used for the purpose are shown in the table below:

Header	**Definition**
title	Title
image_url	Link to the image of the episode
video_url	Link to the video of the episode

In order to provide the user with more detail about the selected episode, we accessed a gigantic 38 MB file (simpsons_script_lines.csv) containing the scripts of all 500+ episodes. Within simpsons_script_lines.csv, each row contains information about a line -- the episode it was in, whether it was spoken, the character who spoke it, the location in which it was spoken, and more. The variables we used are shown in the table below:

Header	Definition
episode_id	Episode ID in which the line was spoken (number)
raw_character_text	The name of the character who spoke the line (String)
raw_location_text	The location in which the line took place (String)

There were several challenges to parsing this dataset into usable data that could be used to analyze each episode. All the code for this parsing is included in the file episode_data.js.

Unlike the first dataset, this one did not include the season number and “number in season” for each episode. Therefore, we had to use the previous dataset to map each episode ID to a season number and an episode number. We did this by creating an array with the max episode ID of each season, and using that to map each episode in the larger database to a season.

Once the first data simpsons_episodes.csv was imported, we used it to generate a 2D array of 27 seasons, each containing a JSON object for each of its episodes. Then, we imported the script lines, and based on the episode_id associated with them, used them to update the JSON for each episode.

This involved updating the total number of speaking lines in the episode, accumulating which characters had speaking lines and how many times they spoke, and accumulating the locations at which speaking lines took place and how many lines were spoken at each location. We only considered speaking lines for the purpose of this project, since they are the most relevant in terms of defining episode quality and rating.

Finally, when we had a JSON for each episode with data about characters and locations, we used some online examples to write a couple of functions that would:

1. Sort each JSON in descending order for bar graph purposes
2. Convert each JSON to a CSV string
3. Download each CSV (two per episode, one for characters, one for locations)

We ended up with a custom database of 1000+ small CSV files that we could import on request when the user clicked the “previous episode” or “next episode” buttons. The imported data was used to generate two bar graphs for each episode. The first bar graph shows the characters (limited to the top 25) in the episode and the lines spoken by each. The second bar graph shows the locations (limited to the top 25) in the episode, and the lines spoken at each location.

Visual Elements

A description of the mapping from data to visual elements. Describe the scales you used, such as position, color, or shape. Mention any transformations you performed, such as log scales.

Throughout the visualization, we have kept a similar color scheme in order to make it visually appealing for the user. All non-black and non-white colors used in the project are from the official Simpsons pantone colors. Following are specific descriptions of our three graphs.

Element 1

Underneath our first plot, we made a slider that user can move left or right to pick a season. The plot interacts with this element, and the title of the plot changes appropriately: for instance, if the user fixes the slider at season 7, the title showing up is season 7: Episode Rating and Viewership. The default season number is 1. The x-axis in this plot is episode number, of which range is calculated by d3.extent function. For each episode, we can access its IMDb rating (vertical axis on the left) and US viewership (vertical axis on the right), and we captured all those points on the plot and connected them as paths in red and blue, respectively. We padded the scales of both y-axes by small amounts to ensure that are interpolation (used to make the lines smooth) did not spill over beyond the x axis.

Element 2

The slider also interacts with our second visual element. Underneath the slider, we made a svg that contains an image and a video url (hyperlinked as “Watch this episode”) which correspond to the current episode the user is looking at. This is done by referencing to the image_url and video_url variables in the first dataset. On each side of the image, there is a button that enables users to move the episode back and forth within the same season by clicking. The default episode number is 1.

Element 3

Lastly, there are two plots at the very bottom of our web page. These plots contain the characters/locations who spoke the most lines/where the most lines were spoken. We limited both to the top 25 for cases where there were more than 25 characters/locations. Both visualizations are bar charts where each bar corresponds to a character/location. The x-axis represents the number of times a character spoke or the number of times a line was spoken at a location. The y-axis represents the characters or locations. Both scales are linear.