Seth Hanusik

Visual Interfaces Project 3: TV Show Scripts

House Scripts

Contents

[Data 1](#_Toc101448125)

[Visualization Components 1](#_Toc101448126)

[Discoveries 2](#_Toc101448127)

[Process 3](#_Toc101448128)

# Data

The data was obtained from [here](https://transcripts.foreverdreaming.org/viewforum.php?f=890), and was scraped from each respective episode’s page using the python library, BeautifulSoup. After scraping, the data was organized into four columns, Character, Line, Season, and Episode, and inserted into a row, which was inserted into a CSV file. This was the data used to generate the three bar graphs and the timeline. The data was processed further using the python library, textcleaner, where stopwords and punctuation was removed from each line before again being inserted into a CSV, and this was the data that was used in the generation of the word cloud. There were six episodes excluded from the data set, S1E2, S1E3, S1E4, S1E5, S7E18, and S8E23 because the scripts were not formatted correctly on the web page, meaning they did not contain the character names of who said the line. S1E9 also shows a significantly lower number of lines, and this is because the site did not contain the full script for that episode.

The characters that were used in the bar charts, were characters that had more than 250 lines of dialogue throughout the entire show. This included the main and recurring characters throughout the series but excluded characters that were only introduced for one episode, like the patient that was being diagnosed by House and his team that episode.

The only words that were used in the word cloud were words that were spoken a specific number of times for the designated character. Dr. House had the greatest number of lines so his threshold for times a word needed to be used is set to 100, which is higher than the other characters, which had their threshold set to 25 uses.

# Visualization Components

Starting at the top of the page, you will see three bar charts, and going from left to right, they represent the total number of lines each character has throughout the entire show, the number of episodes each character appears in, and the number of lines in a selected episode. You can hover over each bar to display a tooltip containing the character and the number each bar represents. For the third bar chart, it starts empty because it is based on a specific episode, and there is a drop-down box to the right of it that the user can use to pick the desired episode.

Moving down to the next row, there is a timeline that spans the entire show. The timeline displays the number of lines per episode of the selected character, and it defaults to House. The selected character can be changed with either the drop-down box on the right side, or by clicking one of the bars in the first two bar charts above. The user can hover over each bar to display a tooltip containing the episode and the number of lines for that episode. The user can also click on the bar to change the episode selection of the third bar chart above it, this will also update the selected episode in the drop-down box and highlight the selected episode orange on the timeline.

The last visualization is the word cloud below the timeline. This displays the most common words spoken by the selected character, and those thresholds were discussed previously. As previously mentioned, the default selected character is House, so the word cloud also defaults to House, and will update when the character selection changes.

# Discoveries

The biggest factor that my application allows a user to discover is how important a character is to the show. With the charts showing the total number of lines and total number of episodes, you can see if there is a correlation between the two. Those two also pair well with the timeline because it expands the detail of the first two charts, and using the timeline to select an episode, allows the user to see if the character has less lines in an episode, who is filling the gap.

One of the things that I find interesting is that while Wilson, Chase, Foreman, and Cuddy all appear in nearly as many episodes as House, he has three to five times the number of lines of dialogue throughout the show (Figure 1). I also like how the timeline gives an insight into when characters are introduced, leave, or start to become less prominent characters. For example, Cameron is one of the main characters through the first three season, but then becomes significantly less important starting with season four, before almost leaving completely part way through season six (Figure 2). The last thing that I find interesting, is that using the word cloud, ‘house’ is always one of, if not the top word for each of the characters, even House himself (Figure 3).

Chart, histogram

Description automatically generated

Figure

Chart

Description automatically generated

Figure

Text

Description automatically generated with low confidence

Figure : Foreman’s Word Cloud

# Process

Starting with the data processing, I used the python BeautifulSoup and textclean libraries to scrape my data and then clean it up for the word cloud. For the application, I used the standard d3 library for creating the bar charts and the d3-cloud library to create the word cloud.

The code was structured so that each chart had its own class and was instantiated from the main.js file. Within the main file, the data was read in from the CSV files and then properly formatted to be inserted into their respective classes. There are also two helper functions that aid in preparing the data format for the timeline and producing the word list of words that reached the desired threshold for the word cloud.

To use the application, the repository can either be cloned from [here](https://github.com/Kisunah/visual-interfaces/tree/main/Project3) and hosted locally using a simple python server, or it can be found hosted publicly [here](https://homepages.uc.edu/~hanusisw/visual-interfaces-project3/). Lastly, you can find a short demo video that gives an overview of how the application works [here](https://youtu.be/bpteRW0bjR0).