RISE RESEARCH INNOVATION SCHOLARSHIP ENTREPRENEURSHIP

Mentor Name. : John Rachlin
Subject/Category : CODE BLUE: NLP Library for Medical Drama TV Shows
Class : DS 3500, HW 7 mini-project
Group # : Kaydence Lin, Tanishi Datta

Northeastern University

CODE BLUE: NLP Library for Medical Drama TV Shows

Kaydence Lin, Tanishi Datta

Background, Motivation, and Goals

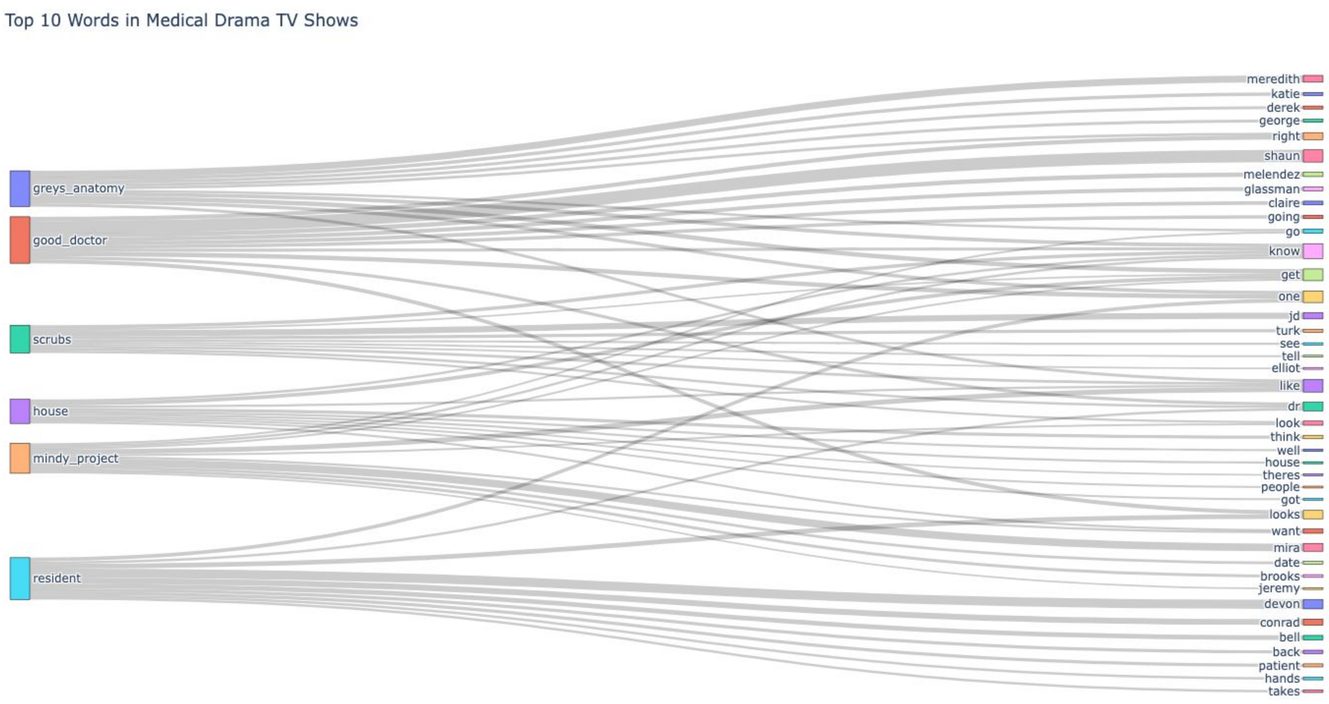
This project aims to conduct an in-depth analysis of scripts from several renowned medical TV series, including "Grey's Anatomy," "The Good Doctor," "Scrubs," "House," "The Mindy Project," and "The Resident." Our objective is to meticulously examine the dialogue, themes, and emotional nuances present within each show. Through this analysis, we seek to gain insights into the characters' development, narrative structures, and the portrayal of medical scenarios. By employing sentiment analysis techniques and identifying recurring language patterns, we aim to uncover the distinctive elements that contribute to the appeal and impact of each series. Additionally, by comparing scripts and calculating cosine similarity, we endeavor to discern commonalities and differences in storytelling approaches and character portrayals across these iconic medical dramas. Ultimately, our goal is to provide a comprehensive understanding of the cultural significance and portrayal of healthcare and medicine within the realm of popular media.

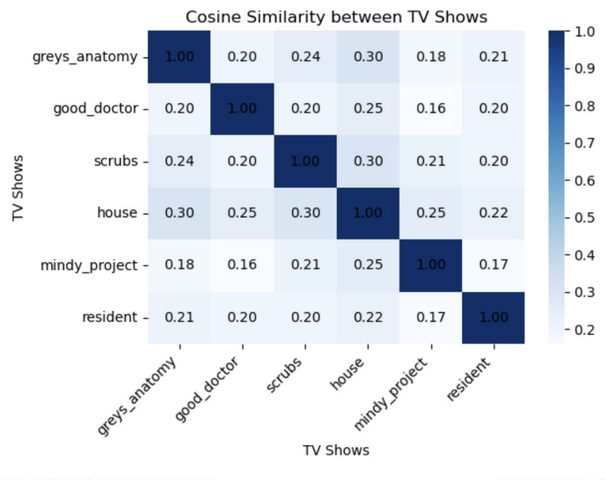
Process and Methods

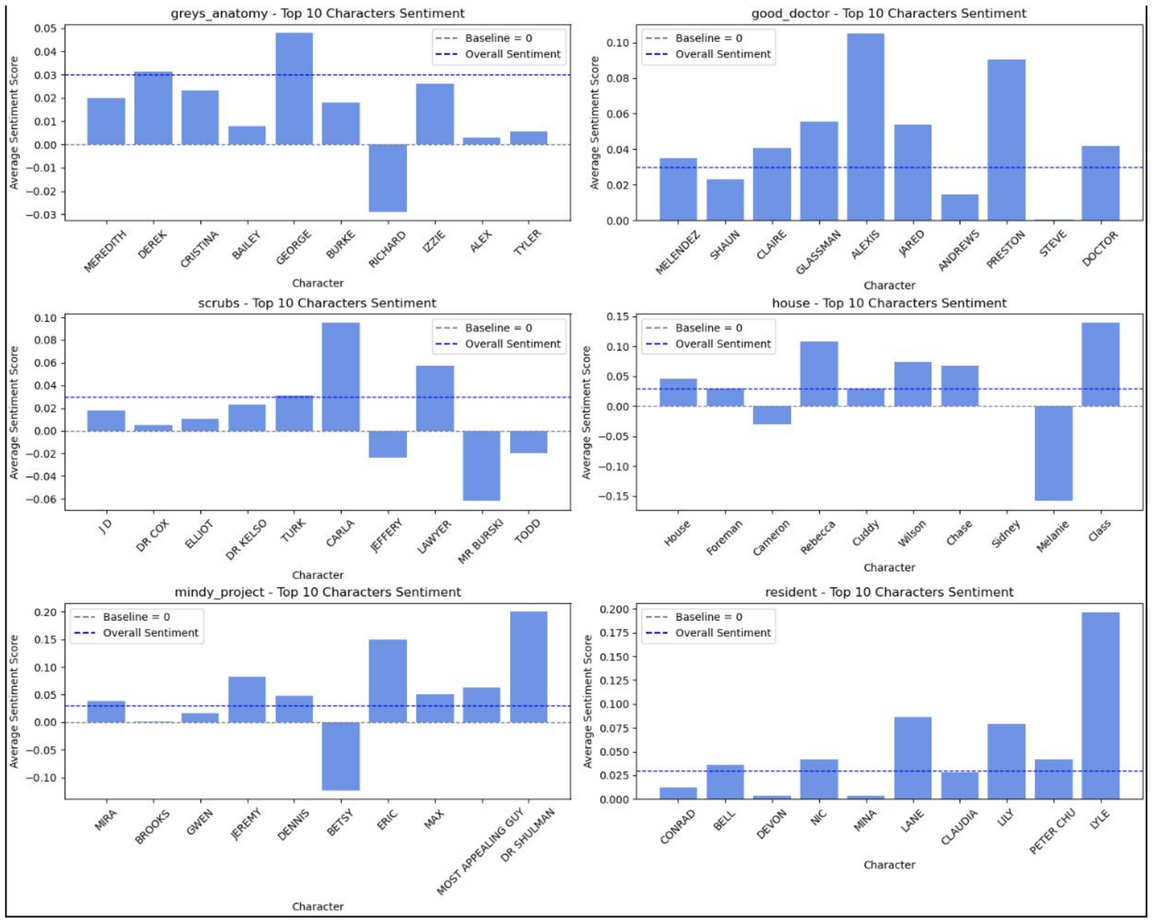
Data collection: We found PDFs of most pilot scripts online and found the House pilot script on a website and put it in a text file. We converted the PDFs to text files in Python.
Script Parsing: We found key words that had non-dialogue items (ex. cues, descriptions, settings) and removed it. We used if statements to clean any typos and non-dialogue words or lines unique to a TV show. We created a dictionary where the key is the character's name and the values are their lines and another dictionary where the word count was stored.
Stop Words: We loaded stop words from a text file to use for cleaning for analysis from the dictionaries.

Findings and Products

Sankey: This Sankey visualizes the most frequent words from the scripts of 6 different medical TV shows. The data suggests a potential overlap in the language used across these medical dramas and comedies, perhaps pointing to commonalities within the genre or main characters in the show.
Sentiment Analysis: Next, we created bar graphs for each show as subplots, depicting the average sentiment score for the top 10 characters in the same set of TV shows. These graphs illustrate the overall sentiment trend within each show's script.
Cosine Similarity: The last graph is the cosine similarity between the dialogues of different TV shows. "Grey's Anatomy" and "House" show the highest similarity score, suggesting the emotion of their dialogues are most alike, while "The Mindy Project" often has the lowest similarity scores with other shows, indicating a unique script style.







Conclusion and Next Steps

Overall, our findings focus on the language and tone of 6 iconic medical TV shows which allowed us to visualize recurring themes and emotion. Understanding these patterns not only provides insights into the storytelling techniques within the genre but also offers valuable information for content creators and researchers interested in the portrayal of healthcare professionals in popular media. A limitation of this project is that we were not able to completely clean and parse the script of non-dialogue items. Even though we were able to identify patterns that removed majority of non-dialogue items, there are still some non-dialogue lines present in the dictionary. These lines are more difficult to clean because someone are in between spoken dialogue with no identifying characters or patterns. However, these lines should make up a minority in the dictionary compared to a character's spoken dialogue. Next steps that we could take is possibly non-dialogue lines to see if settings, descriptions, and directions are similar. We could also do an in depth analysis of each character from different shows depending on specialty and position, like nurses, surgeons, and anesthesiologist. It will also be interesting to look at n-grams, like bi-grams, to see which phrases or medical terminology is used often. We can also use it to analyze if a show is medically accurate to prevent misinformation to viewers.
Authors Contributions: Kaydence Lin - Parsing text files and Cosine Similarity; Tanishi Datta - Sentiment Analysis and Sankey Diagram