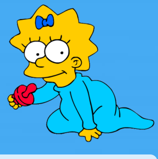
Gender in THE SIMPSONS



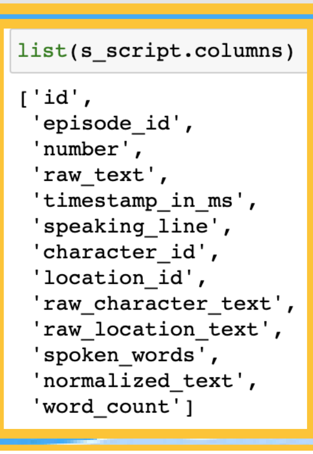
Justin Carlou "Jus" Lim | DigHum100 Dr. Adam Anderson | Summer 2021

Project Description

For this project, I will be exploring the representation of gender in the television show, *The Simpson's*.

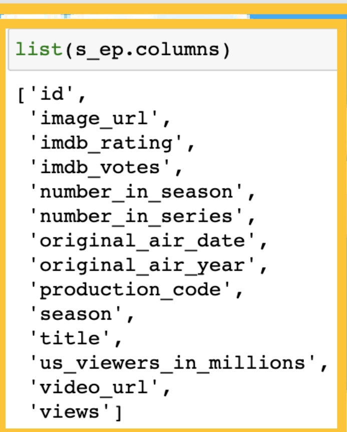
Dataset

In this project, I use Prashant Banerjee's "*The Simpson's* Dataset" posted on Kaggle. This dataset includes information about each line a character says in the show, where the characters say the line, and IMDb ratings for each episode. The dataset is split into four datasets: the character dataset, the episode dataset, the locations dataset, and the script lines dataset. Each were combined and formatted using the pandas library. The column names of each data set are listed below:



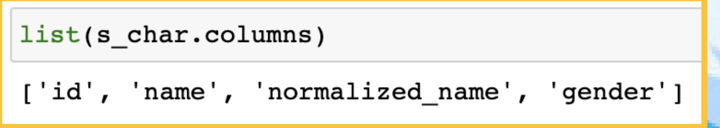
```
list(s_script.columns)

['id',
 'episode_id',
 'number',
 'raw_text',
 'timestamp_in_ms',
 'speaking_line',
 'character_id',
 'location_id',
 'raw_character_text',
 'raw_location_text',
 'spoken_words',
 'normalized_text',
 'word_count']
```



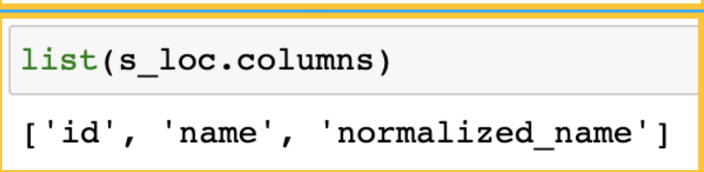
```
list(s_ep.columns)

['id',
 'image_url',
 'imdb_rating',
 'imdb_votes',
 'number_in_season',
 'number_in_series',
 'original_air_date',
 'original_air_year',
 'production_code',
 'season',
 'title',
 'us_viewers_in_millions',
 'video_url',
 'views']
```



```
list(s_char.columns)

['id', 'name', 'normalized_name', 'gender']
```



```
list(s_loc.columns)

['id', 'name', 'normalized_name']
```

Questions for Analysis

1. How has the inclusion of female characters changed over time?
2. How does dialogue between male characters differ from dialogue between female characters?
3. How often does *The Simpsons* pass the Bechdel Test?

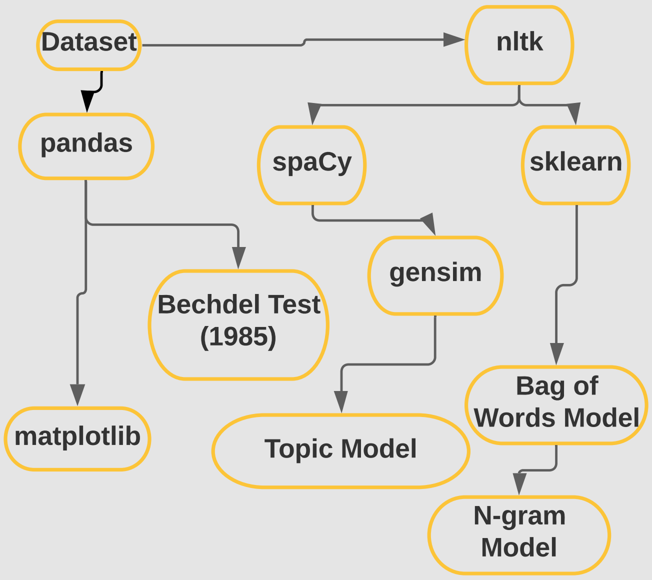
Hypotheses

1. Because *The Simpson's* is a show which follows pop culture closely, I do believe that the preportion of female characters will change, but only slightly.
2. I also believe that conversations between folks of the same sex will be stereotypically gendered.
3. Yet, I imagine *The Simpson's* will often pass the Bechdel Test.

Notes

** expressions of gender in this project are confined to a binary because that is what is provided in the dataset. Further analysis in including nonbinary genders would be fruitful to explore in the future.

Tools & Methods



Dataset

nltk

pandas

spaCy

sklearn

gensim

Bechdel Test (1985)

matplotlib

Topic Model

Bag of Words Model

N-gram Model

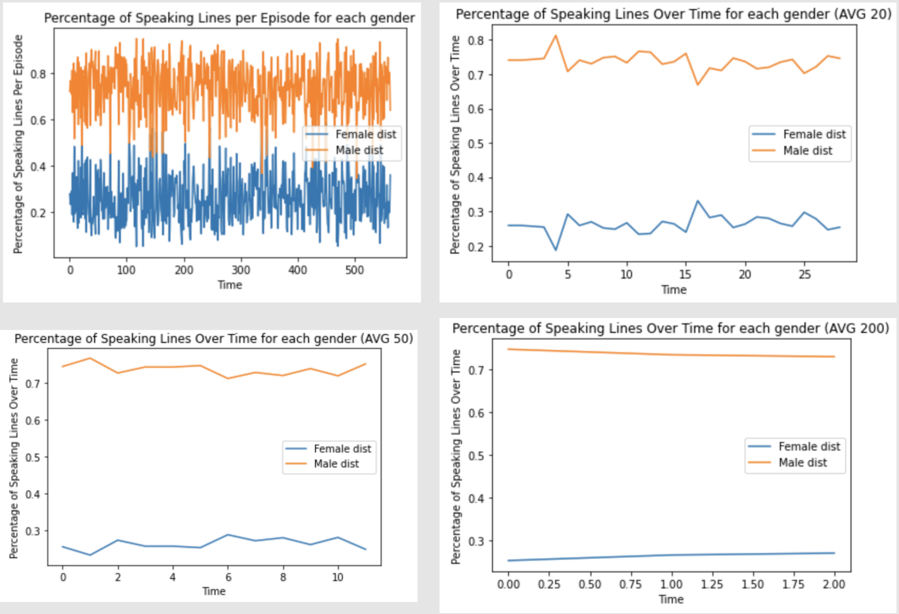
Workflow

In order to track representation of women over time, I used methods from the pandas library to find the distribution female speaking roles to male speaking roles over time, then used matplotlib to visualize the results.

To compare dialogue between male and female characters, I created a topic model out of each group's dialogue, then visualized the topic network. Moreover, I supplemented the analysis with an N-gram model to see differences in actual phrases.

Lastly, I used pandas to sort through character dialogue for each episode, and created a function to test the Bechdel Test.

Gender Distribution Over Time



I plotted the gender distribution of speaking roles for each episode in the top left graph. As it resulted in overplotting, I took the average of an increasing group of data points (from 20 - 200) to show clearer trends. As we can see from the bottom right graph, there seems to be an increase in female speaking roles over time, but it is small and insignificant.

Works cited:

- https://www.kaggle.com/ruchi798/sentiment-analysis-the-simpsons/notebook#Unigrams,-Bigrams-and-Trigrams
- https://www.kaggle.com/prashant111/the-simpsons-dataset
- https://towardsdatascience.com/topic-modelling-in-python-with-nltk-and-gensim-4ef03213cd21