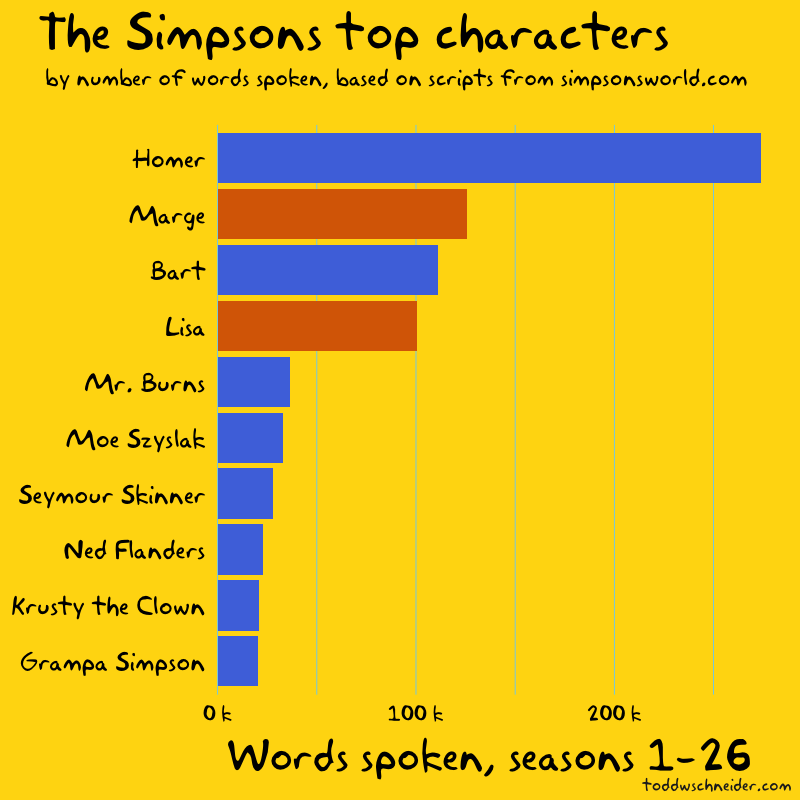
Aim

Branching off of the work done by Todd Schneider where he uncovered that out of the top ten characters, according to number of words spoken, only two of those were women. This got me thinking about gender disparity within the show. I want to look into gender representation on the show and see if I am able to predict it based on character language. Using natural language processing I will analyze raw text of spoken lines by character and create a model that tries to predict gender based on variables pertaining to language. My goal is to accurately predict gender from language and further our understanding of predictive learning.

I came across *The Simpsons by the Data*, by [Todd W. Schneider](http://toddwschneider.com/), where he used data collected from a Simpson’s database to gain insights on the significance of secondary characters, the decline in TV ratings, and uncovered unequal representation of gender between characters.



Gender imbalance on The Simpsons

The colors of the bars in the above graphs represent gender: blue for male characters, red for female. If we look at the supporting cast, the 14 most prominent characters are all male before we get to the first woman, Mrs. Krabappel, and only 5 of the top 50 supporting cast members are women.

Women account for 25% of the dialogue on The Simpsons, including Marge and Lisa, two of the show’s main characters. If we remove the Simpson nuclear family, things look even more lopsided: women account for less than 10% of the supporting cast’s dialogue.

A look at the show’s [list of writers](https://en.wikipedia.org/wiki/List_of_The_Simpsons_writers) reveals that 9 of the top 10 writers are male. I did not collect data on which writers wrote which episodes, but it would make for an interesting follow-up to see if the episodes written by women have a more equal distribution of dialogue between male and female characters.

In Delving deeper into the work done by Schneider on the subject of gender representation on the show, if successful in predicting gender using machine learning, I would be interested in further exploration with the information gained from an accurate model.

Methods and Models

The method I will be using will be natural language processing (NLP). Within NLP I will be utilizing the natural language toolkit (NLTK). The model I will be using will be naïve Bayes classifier.

The type of naïve Bayes I will most likely be Bernoulli due to the reasoning that it utilizes word occurrence vectors as opposed to word count vectors.

I have not decided yet whether or not I will need to run Out of Core model fitting to reduce the computational overload a large dataset imposes. If the dataset is too large I will do a partial fit on the data with the maximum size data chunks that the RAM allows.

Risks & Assumptions

In setting gender to my y-variable I am assuming that language in itself is a predictive indicator of gender. This is going off of the idea that women are represented in specific roles relating to recognizable words or phrases.

Goals & Success Criteria

My goal is to create and train a model that accurately predicts a character’s gender given normalized text data. The model will be deemed successful if it can correctly classify gender at or above a model score of 0.85.

Data Cleaning/Munging Techniques

I initially uploaded two datasets. The data frame, df\_3 contains the text spoken during each episode (including details about which character said it and where) along with the character who spoke. The raw text is also normalized and within a separate column labeled as such. Out of that dataset there were 23 rows containing errors that could not be imported due to inconsistent numbers of columns. I decided to drop the rows due to the overall large size of the data (length is 158315).

Of that data null values for character id and text is 17,521 and 17,522 out of 140,7726. Nulls for this data frame will be dropped, again because of overall data size I do not deem the dropped rows significant.

The df\_4 data frame contains character id, normalized name, and gender. This data frame contains the information that will be used as my predictive y-variable. Analysis of this data frame showed a length of 6,722 rows, which contained 6,399 null values for gender. At first glance this was a major roadblock for me, but upon further investigation I discovered that all the null values can be dropped. I found an alternate version of this data frame and imported it as df\_5. The alternate version contains the 323 non null rows. In the work done by Todd Schneider he created the data frame based on the top 323 characters, which accounts for 86% of the show’s dialogue.