The Simpsons: Predicting episode ratings using episode scripts.

By Mandi Uppal

Purpose

- The aim of this project is to utilize scripts from the popular show “The Simpsons” to predict episode ratings.

- These findings can be used by busy networks bombarded with pilot scripts to save time and money on the selection process, filtering through less promising show ideas.

Data

- The datasets for this project were obtained from the website *data.world* via the link: https://data.world/data-society/the-simpsons-by-the-data.

- In order to access the datasets a free account was created in order to join the website as a member. The *simpsons_script_lines.csv* file containing raw text scripts per episode throughout "The Simpsons" series, as well as normalized text and and spoken word text was downloaded. Additionally, the *simpsons_episodes.csv* file was downloaded, containing imdb ratings for each episode in the series.

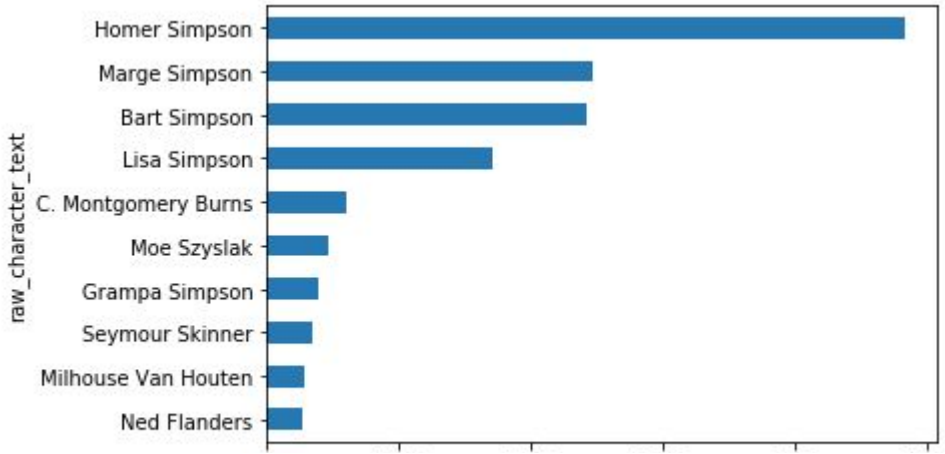
- These csv files were imported into python where they were joined to create one dataset in the format of a dataframe. All irrelevant columns were dropped leaving *imdb_rating* column as well as the *raw_text* column. At this time the extent of data cleaning involved removing rows with missing values and transforming the *raw_text* column of the dataframe into an array.

Exploratory Data Analysis

- Next steps involved exploratory data analysis including:
  - exploring the frequency of lines per character over the course of the series
  - length of lines per character over the course of the series
  - the distribution of episode ratings for the series
  - the distribution of the number of views for the series

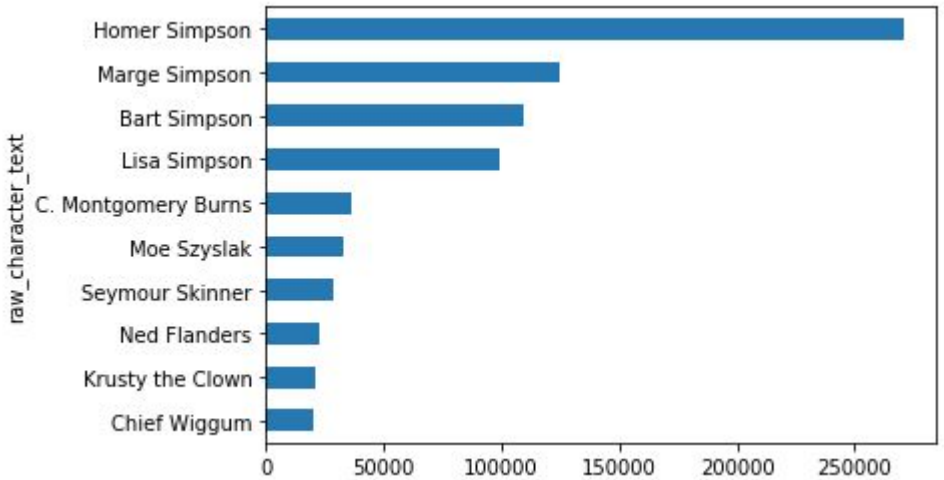
Number of Lines Per Character

We can see that the characters in the plot above are the top ten characters with the most lines throughout the series.



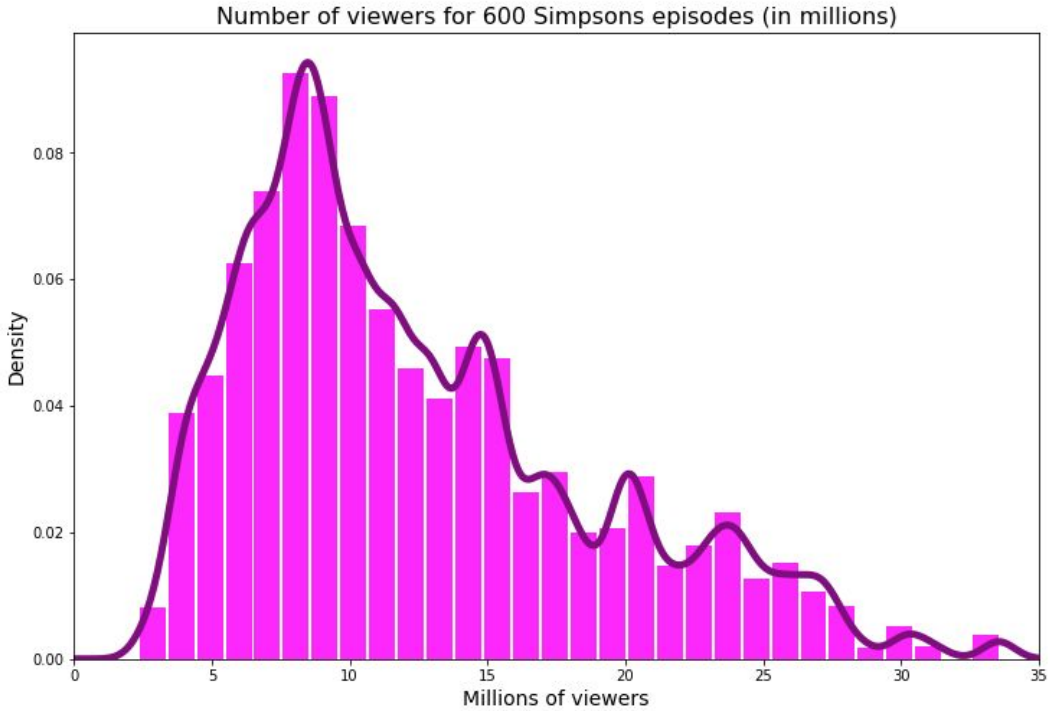
Number of words per Character

The number of words per character that are spoken throughout the series follows a similar pattern to the number of lines that a character speaks within the top ten group.



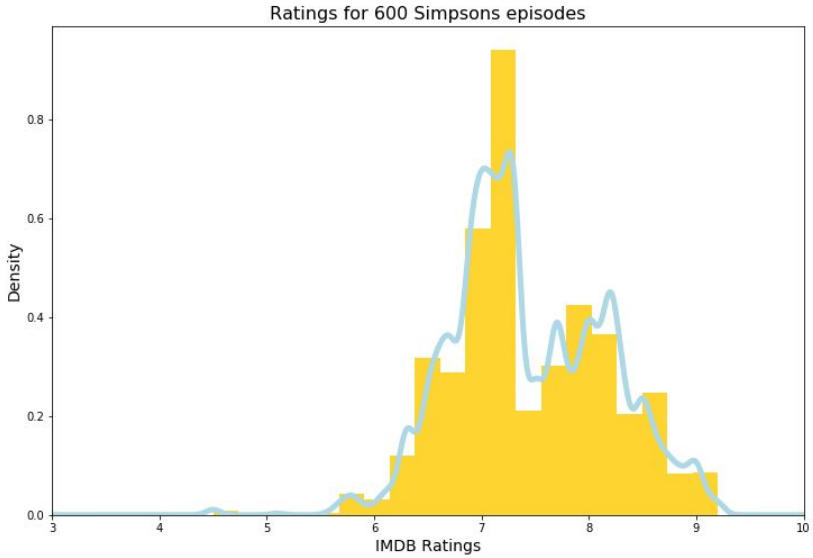
Number of Views

We see here that the number of viewers per episode is mostly 8 or 9 million views but has been as high as over 30 million views.



Distribution of Ratings

Episode ratings seem to be most frequently rated around 7.0 out of 10.



Machine Learning

- The models selected for this classification prediction problem of predicting show rating were ***logistic regression*** and ***random forest***.

- Data pre-processing included the *tokenization or tagging* of the script data from the *raw_text* column. Next we trained a *Doc2Vec* model using the tagged data, the resulting vectorized text was utilized in the training of our logistic regression and random forest models.

Conclusion and Next Steps

Results:

- The results indicate that both logistic regression and random forest models performed well, overfitting on the training set (1.0 accuracy) and 96% accuracy on the test set for ***random forest***.

- Our ***logistic regression*** model performed similarly with an accuracy of 95% for both the train and test sets.

Next Steps

- Next steps should include **further refinement of models** involving:
    - *Hyperparameter tuning*
    - Utilising the results of *exploratory data analysis* as well as *statistical data analysis* to inform which features might enhance model performance and thus prediction.
    - For example, *feature enhancement* might include exploring whether the gender of a character has an impact on the number and length of lines that they have.
    - Additionally, we could experiment with alternative *text vectorization* methods, such as bag-of-words.