Lord of The Rings: Who's speaking?

By: Yaniv Bronshtein
Link:
https://www.kaggle.com/paultimothymooney/lord-of-the-rings-data

Data Description:
The Lord of the Rings Character Descriptions and Dialogue Data Set is extracted from the iconic *Lord of the Rings* trilogy based on the book series by J.R.R Tolkien. This Dataset includes two .csv files: 'lotr_characters.csv' and 'lotr_scripts.csv'. The goal is to explore both data sets to gain statistical insights and perhaps explore the potential of converting the problem from Data Analysis to one of Machine Learning.

Let us first discuss 'lotr_characters.csv'. We see five columns: 'birth', 'death', 'gender', 'hair', and 'height'. Every column is of type string, so it will be a challenge to convert the columns into useful numeric features. It is suspected that more complicated feature engineering and preprocessing will be required.

Now let us discuss lotr_scripts.csv. This csv files contains only three columns: 'char', 'dialog', and 'movie'. Just as in the previous table, the columns are all of type string. Perhaps joining the two tables will provide more insights as opposed to treating the two tables as individual entities.

At the end of the analysis(long-term), one should be able to answer the following questions:

1. Given a piece of text from any one of the three movies, who said it and what other features correlate to the classification?
2. What features are most correlated?
3. Does the data reveal any bias?
4. How does a character grow change throughout the trilogy?

It is a conscious decision to focus on the existing LOTR trilogy instead of getting more data from Harry Potter. The data cleaning process is more challenging but I believe will be more rewarding. The characters in Harry Potter are more homogenous while Lord of the Rings contains more races and mixtures and even characters born in different ages.