Data Collection

By Jim Morgan, Foad Namjoo, Ariel Oconitrillo

Data Overview

Our origin data consists of a csv download from kaggle.com comprising 36 out of 38 plays written by William Shakespeare. Additional information about the plays was sourced from the Royal Shakespeare Company's website.

Data Structure

The data is split into six columns, Dataline, Play, PlayerLinenumber, ActSceneLine, Player, and PlayerLine. These columns will allow us to perform hierarchical analysis of the different characters in each of Shakespeare's plays. The dataset contains 111,396 lines as indicated by the Dataline column. The ActSceneLine contains the smallest number of actual data points, with null values balancing out the content. These null values are useful indicators for where the PlayerLine delivers stage direction as opposed to useful dialogue. Actual character lines amount to 105,153 lines across the 36 plays.

We plan on storing the data intact in CSV format, and then manipulating the data in-code to a pandas dataframe. The source dataset is relatively clearn, requiring minimal processing to elminate irrelevent features, like stage directions.

Data Analysis and Processing

Part of data analysis will be through characterizing the data in terms of the source play and the character speaking the lines. This characterization will also include histograms of word counts and usage, n-gram analysis, K-mean clustering and TF-IDF. By using K-Means and n-grams, we compare the plays with each other, and with the additional film transcripts described below. Our clustering algorithm will use the "elbow curve" technique to determine cluster size. Our n-grams include 2, 3, and 4 word lengths.

Additional Data

Additional data sets yet to be fully characterized include several movie transcripts for The Lion King, She's the Man, 10 Things I Hate About You, and Warm Bodies. These films are either directly or loosely based upon Shakespeare's works, and performing analysis on these films is part of the scope of the initially proposed analysis. Further work is being done to refine the data into a usable source