Data Appendix

“Theoffice” Dataset

We did not perform analysis on this dataset, but we modified it and used it to adapt our other dataset. We used the other dataset for our final analysis. Each observation in this dataset represents a single line in a script from an episode of The Office. The dataset includes the script of every episode from every season of The Office. There are hundreds of lines in each episode so this dataset is very large. The variables in this dataset are listed below.

season

The “season” variable states the season that this episode aired during. It is 1-9, but we chose to only include season 1-6, since Michael leaves the show and we wanted to be consistent.

episode

The “episode” variable states which episode each line in the script is from. It is between 1-24.

episode_name

The “episode_name” variable states the same as the “episode” variable, but includes the name of the episode. It varies based on the episode, it is never the same.

director

This is the name of the person who directed the episode. It varies based on the episode and season.

writer

This is the name(s) of the person/people that wrote this episode. It varies based on the episode and season.

character

This is the person that said the line. It varies based on script. We are primarily focused on Michael, Dwight, Jim, and Pam. We used this to subset this dataset and create our second dataset.

text

This is what the character said.

text_with_direction

This is what the character was supposed to be doing when they said their line.

imdb_rating

This is the rating of the episode from imdb. It is 1-10.

total_votes

This is the number of votes that were taken into account for the imdb_rating.

air_date

This is the day/month/year that the episode aired.

“new_df” Dataset

This is the only dataset that we used for our analysis. There were only two variables in the dataset, but it changed slightly for each character that we analyzed. This dataset was created by subsetting the first data set to only include one character at a time. The characters were Michael, Dwight, Jim, and Pam. We also subset the dataset to only include seasons 1-6. Then we had the dataset count the number of lines for the chosen character per episode. Each observation in this dataset is the name of the episode and how many lines a specific character spoke during that episode. It varies from Michael, Dwight, Jim, and Pam since we performed this analysis on all four of them.

lines_per_episode

This is the number of lines that the character had per episode. Since we have a loop that we changed for each character (Michael, Dwight, Jim, Pam), there are 4 different variations of this.

imdb_rating

This is the rating of the episode. It does not change based on the character that we are using.