Sam Bacon
CSC401
May 2, 2021

Final Project: Proposal (Task 1)

Problem/Questions:

As stated in Lab 17, I have decided to do a thorough analysis of the scripts of all lines of *The Office*. Movies, TV shows, and a variety of streaming services command a massive share in the global entertainment industry. If a new show is released and its popularity explodes, there is no limit on how successful and profitable that series can be. However, producers and film crews are operating in a highly competitive market, with dozens of new movies and shows released every week. Bearing the cutthroat nature of this industry in mind, it is crucial that creators understand what makes a show or movie successful. How long are the episodes/seasons? What are the character dynamics? How long are the scenes? The better producers understand the intricacies of a successful movie or series, the better off they will be as they look to create the next blockbuster.

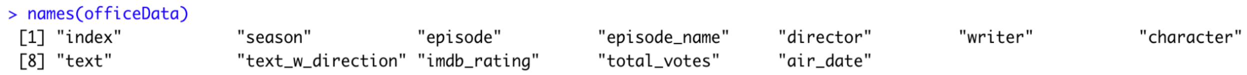
The Office (US Version) is a wildly popular TV show that ran from 2005 until 2013. The writers and cast members collected countless awards, and it seems that everyone you ask has seen at least a few episodes (or every episode a few times). If a producer hopes to create a show that will captivate audiences for years to come, *The Office* serves as a golden example. This analysis will seek to gain a better understanding of how *The Office* is structured. Specifically, I will be examining the following topics:

1. Sentiment: What is the overall sentiment/polarity? Does this change based on characters? Does it change based on season or episode?
2. Fluidity: How long are the lines? Again, how might this change based on character or episode?
3. Ratings: Is there any relationship between the two factors described above and the IMDB rating for that specific episode/season? Can we create an algorithm to model this relationship?

The answers to these questions may not be straightforward. I anticipate to have some difficulty cleaning and formatting my dataset. There may also be challenges with determining what variables have a causal relationship and which variables are simply associated. I anticipate that I will observe differences in the sentiments based on character, but I am not sure whether sentiment will also change based on episode or season. I expect the length of the lines to vary based on the episode, but it will be interesting to see if they also change based on character. Regarding ratings, I expect to see some relationship between attributes and the IMDB ratings (see methods section).

Data:
I will be using data from the ‘Schrute’ R package.
(https://technistema.com/posts/introducing-the-schrute-package-the-entire-transcripts-from-the-office/)

This dataset contains every line spoken throughout the nine seasons of *The Office*. There are 55,130 total entries, each with the following attributes.



```
> names(officeData)
 [1] "index"            "season"           "episode"          "episode_name"     "director"         "writer"           "character"
 [8] "text"             "text_w_direction" "imdb_rating"      "total_votes"      "air_date"
```

I will store this dataframe in Google Sheets using the googlesheets4 package in R. Then, I will download the document, so the full dataset will be stored on my desktop. From there, I can easily import the data into python, where I will perform my analyses.

Methods:

1. Sentiment analysis and length:

   I will preprocess the “text” and “text_w_direction” attributes using the same steps that I used in Homework 4: Missing Topic. This consisted of sentence and word tokenization and removing stopwords using the NLTK package. After that, I will also be able to perform word stemming, lemmatization, an part-or-speech tagging to create some preliminary visualizations. Finally, I will use the TextBlob package to analyze the polarity and subjectivity/objectivity values of each line. With these two new attributes, I plan to create a K-means clustering model to group the lines based on combinations of length, polarity, and subjectivity/objectivity. It might also make sense to obtain the TF-IDF score of each line and create a Naive Bayes classification model to predict sentiment.
2. Ratings:

   The ultimate goal of this analysis is to provide insight as to what factors are needed to create a successful show. I plan to train a variety of models and see how accurately they can predict the IMDB rating of the episode. I anticipate performing some principal component analysis to hone in on the most important variables (not sure what they will be yet). From there, I will create a KNN model to see if ratings can be predicted based on certain combinations of attributes.

Evaluating Success: I would like to be able to provide 2-3 meaningful insights from these models. That does not necessarily mean that these models have to have a high accuracy. If none of the models ended up being very accurate, that is still valuable information that is worth sharing. My report will be successful when it has thoroughly investigated a variety of possible relationships and models, which I intend to do.

Visualizations:

1. Sentiment Analysis:
    a. *Frequency plot of most common words* and part-of-speech (may also block based on character or season)
    b. These visualizations should provide some more preliminary info about the dataset and see if there are any stark differences between the characters that are worth investigating.
2. Ratings:
    a. Scatterplots comparing various quantitative attributes and rating
        i. X-axis: length, polarity, subjectivity/objectivity
        ii. Y-axis: IMDB rating
        iii. Again, these could also be blocked based on character.

Note: overall, I plan on having at least 10 visualizations. These will be very helpful as I look for specific patterns or correlations that are worth further investigation.

Grade Targets:

Low-Target	The final notebook is unorganized and difficult to follow. Visualizations and models are not labeled clearly, nor is there any in-depth analysis. The presentation is unprofessional or too lengthy. It is difficult for the audience and professor to grasp the main findings/takeaways from the project. The project is lacking some elements that are important for a thorough investigation.
Medium-Target	The notebook is organized for the most part and relatively easy to follow. The visualizations and models are labeled correctly, and the commentaries on these components are accurate and precise. The presentation is not too wordy for the most part, and the audience can clearly understand the main findings. Overall, the project meets all the requirements of a sound investigation.
High-Target	All parts of the notebook and presentation are organized and clearly labeled. Comments are concise and insightful, and the various analyses are organized in a logical pattern. Visualizations are aesthetically appealing and augment the report as a whole. The presentation is well-prepared and slides are not wordy. The key findings of the analysis are clearly stated throughout the

	presentation. The project meets all the requirements of a sound investigation and also provides ideas about how the investigation may be expanded in the future.