
# Who's the boss?

## Midpoint Review Deck

# Table of Contents

# Highlights


## Accomplishments during the sprint:

- Downloaded data, conducted exploratory data analysis, pre-processed text data and created a processed data file for model building
- Set up AWS EC2 and RDS instances and S3 buckets
- Wrote scripts for reproducible workflow

# Review Progress

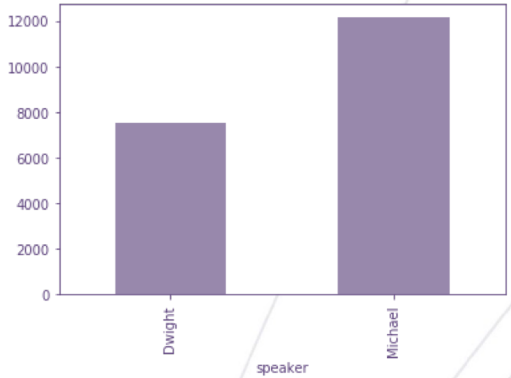

**Epic 1:** Data Ingestion & Understanding: *Preparing the data for future use and cleaning to input into model*

- * Getting the data (0) – Data read from Google Drive and uploaded to public S3 bucket
- * Uploading the data to RDS (1) – Decided against it. Instead using the RDS to save user inputs
- * Data cleaning (2) – Text data had to be pre-processed, stop words had to be removed using pre-existing English stop words
- * Exploratory Data Analysis (4) – Analysed Michael and Dwight's lines from all seasons of the shows to see if they have any peculiarities in the way they speak
- * Feature Engineering (8) – In the process of creating scripts but not complete

# Analysis


*All analysis can be found in the EDA.ipynb. Below are the highlights:*

## Michael vs Dwight (number of lines):




*Despite being on the show for more seasons, Dwight has fewer lines that Michael.*

## Michael vs Dwight (common words):



*The above word clouds display the top 30 words spoken by Michael (left) and Dwight*

# Lessons Learned


- Number of observations for this dataset is limited, especially after we cut it down to only using Michael and Dwight's lines. A one-time trained model would be all we could do with this dataset

- Several pre-existing text analysis packages are limited. For example, when using the nltk stop words list, I found that a lot of common English words weren't included in the list. I ended up adding a few other common words

- In terms of creating a reproducible workflow, I learned the importance of configurations. For example, saving a file to your local system vs an S3 bucket would require very different configurations that could be passed very differently

# Recommendations

The next sprint will primarily focus on model building and creating an interface for the user to interact with

- Model Building. Generate initial models (8)
- Model Building. Evaluate model (1)
- Model Building. Create model pkl (1)
- Application. Build prediction pipeline (4)
- Application. Deploy on EC2 instance (8)
- Application. Build front-end (4)
- Application. Create test cases (2)
- Application. Evaluate user inputs (2)
- Application. Log user input and errors (2)