IST 652

Data Analysis Final Project Report: The Office Dataset

By Jenny Ha & Patrick Walsh

About the Data

In this analysis, the dataset 'the-office_lines.csv' was processed and analyzed. This CSV-formatted data file was obtained from the Kaggle website (https://www.kaggle.com/datasets/fabriziocominetti/the-office-lines?datasetId=1807639) and contains all the dialogue spoken by characters from seasons 1 to 9 of the popular TV show, The Office. Additionally, the dataset also includes dialogue from deleted scenes. The data was acquired through web scraping from the website 'https://www.officequotes.net/' (Cominetti, n.d.). The Office dataset consists of 58,721 rows, each representing an individual line spoken by a character. The dataset is structured into five columns, each providing specific information about the dialogue (line), such as the character that said the line, the season, and the episode number. Below is a table (Table 1.) that presents the column names, definitions, and data types, along with a screenshot (Screenshot 1.) showcasing the actual data. The dataset required minimal preprocessing since it was already well structured, but the program does split the Line column so that the tokens are packed into a list for later processing.

Table 1. The Office Dataset

Column Name	Definition	Data Type
*Unnamed	Row index	Integer
Character	Name of character that said the line (quote)	String
Line	The line spoken by the character	String
Season	Number of the season (1 to 9)	Integer
Episode_Number	The number of episode (varies with the season)	Integer

Screenshot 1.

	Unnamed: 0	Character	Line	Season	Episode_Number
0	0	Michael	All right Jim. Your quarterlies look very goo...	1	1
1	1	Jim	Oh, I told you. I couldn't close it. So…	1	1
2	2	Michael	So you've come to the master for guidance? Is...	1	1
3	3	Jim	Actually, you called me in here, but yeah.	1	1
4	4	Michael	All right. Well, let me show you how it's don...	1	1
...	...	...	...	...	...
58716	61302	Creed	It all seems so very arbitrary. I applied for...	9	23
58717	61303	Meredith	I just feel lucky that I got a chance to shar...	9	23
58718	61304	Phyllis	I'm happy that this was all filmed so I can r...	9	23
58719	61305	Jim	I sold paper at this company for 12 years. My...	9	23
58720	61306	Pam	I thought it was weird when you picked us to ...	9	23

Word Cloud

Method of Analysis

Descriptive statistics were employed as the method of analysis to address both question 1 and question 2. For question 1, the total number of words spoken by each character was computed and subsequently sorted in descending order. The analysis focused on the 'Character' and 'Line' columns. The expected outcome was a list of character names accompanied by their respective total word count. The first row would identify the character that said the most words on the show.

Similarly, question 2 followed a comparable approach. However, in this case, the total number of words spoken by each character was calculated per season. The analysis incorporated the 'Character,' 'Season,' and 'Line' columns. The expected result was a table presenting the character who spoke the most words in each season, along with the corresponding season number and the total word count."

To address question 3, the Naive Bayes machine learning (ML) algorithm was employed to classify the dialogue spoken by the five main characters. The analysis used the 'Character' and 'Line' columns as features for training and testing the ML model. The expected outcome of this analysis includes the accuracy score (%) of the ML model in correctly classifying the dialogue to one of the main characters, as well as a confusion matrix that compares the predicted results to the actual values.

For question 4, the Flair sentiment analysis model was leveraged. The model uses built-in sentiment analysis capabilities to produce a score of positive, neutral, or negative. To determine the overall tone of the show, the Flair model analyzed the script, line by line, reading the vocabulary, word order, phrases, punctuation, and other grammatical markers to infer tone. The model then outputs a score and tone label.

As for question 5, the program used several built-in functions to parse the lines, group the lines together by character, and count the number of lines for each character in the script. From there, the program sorted the counts and displayed the results by highest count and lowest count.

Finally, the last question was addressed through N-gram analysis techniques in Python. The SciKit-Learn library was brought in to count N-grams of 4 and 5, identifying phrases that are 4-5 words in length. This analysis was applied to the main character, Michael, based on his having the largest number of lines in the script. From here, the N-grams were ranked based on the highest count, and the most common phrases were displayed in the results.

Questions to be Answered

1. Which character said the most words in the TV show overall?
2. Which character said the most words per season?
3. Can the main character be determined by the dialogue?
4. What is the overall tone of the TV show in terms of its script?
5. Which characters have longer lines and which characters have shorter lines?
6. What are the top phrases used by the main character?

Program Description

The Office dataset is stored in a CSV format named 'the-office_lines.csv' and is imported into a Pandas dataframe called 'df.' The dataframe contains 58,721 rows and 5 columns. Additionally, the string characters in the 'Line' column are split by empty spaces and stored in a list. A new column called 'Length' is then created to store the count of strings in the 'Line' column.

To address question 1, the 'df' dataframe is grouped by the values in the 'Character' column to calculate the total number of words spoken by each character. The results are stored in a variable called "character_df." This variable is then processed, converted into a dataframe, and sorted in descending order based on the total number of words spoken by each character.

For question 2, the 'df' dataframe is grouped by the values in both the 'Character' and 'Season' columns to determine the total number of words spoken by each character in each season. The results are stored in a variable called 'season_df.' This variable is further processed, converted

into a dataframe, and filtered to extract the characters that have spoken the most words per season.

Regarding question 3, a new Pandas dataframe named 'df2' is created by reading the Office CSV file. The rows in 'df2' are filtered to include only the five main characters on the show: Michael, Dwight, Jim, Pam, and Andy. After filtering, 'df2' consists of 35,062 rows and 5 columns. To train the Naive Bayes algorithm from the Scikit Learn library, the output variable 'y' is initialized with the values from the 'Character' column. Additionally, the input variable 'X' is created using the CountVectorizer() function from Scikit Learn, which transforms the text in the 'Line' column into a matrix of token counts. 'X' has a dimension of 35,062 rows and 16,884 columns representing unique words. Subsequently, the input and output variables, 'X' and 'y,' are split into 80% for training and 20% for testing using the train_test_split() function. The Naive Bayes machine learning algorithm, 'mnb_model,' is then initialized using MultinomialNB() and trained with the training variables. After training, the model is tested with the testing variables to obtain predicted values, and the accuracy of the model is evaluated. Finally, a confusion matrix is plotted to visually compare the predicted results against the actual values, providing insights into how many predicted results were correct and incorrect.

For question 4, the Flair sentiment analysis model was employed. The Flair model's built-in sentiment analysis engine was leveraged to produce a sentiment score as output, taking in the text of the dataset's script as input. The output is a decimal number between 0.0 and 1.0 as well as a label of either 'Positive,' 'Neutral,' or 'Negative.' The closer to 1.0, the more positive the final label output will be.

As for question 5, the program leveraged built-in Pandas functions to find individual characters in the script (using .unique()) and estimated the number of lines each character had with .groupby() and .size() built-in functions. The program then uses sorting functions (using .sort_values()) to get the top 10 and bottom 10 characters with the highest and lowest number of lines. The top 10 and bottom 10 were found using the .head(10) function with a '10' inside the parentheses.

Finally, question 6 employed N-gram analysis to find the top phrases used by Michael, the main character in the script. Michael was determined to be the main character based on having the most dialogue in the script, as seen in question 1. The program uses SciKit-Learn's CountVectorizer() method to do an N-gram analysis of N-grams 4-5 tokens in length. This CountVectorizer() was then run on Michael's lines from the script to find 4 and 5 word phrases, counting up the most common phrases he uses. Surprisingly, Michael's famous "That's what she said!" sits at number 11 in this list, with phrases like "all right all right" and "we are going to" occurring more frequently.

Results

Question 1

The most words said by a character

Character	Total Words
Michael	172,551

Question 2

The most words said by a character per season

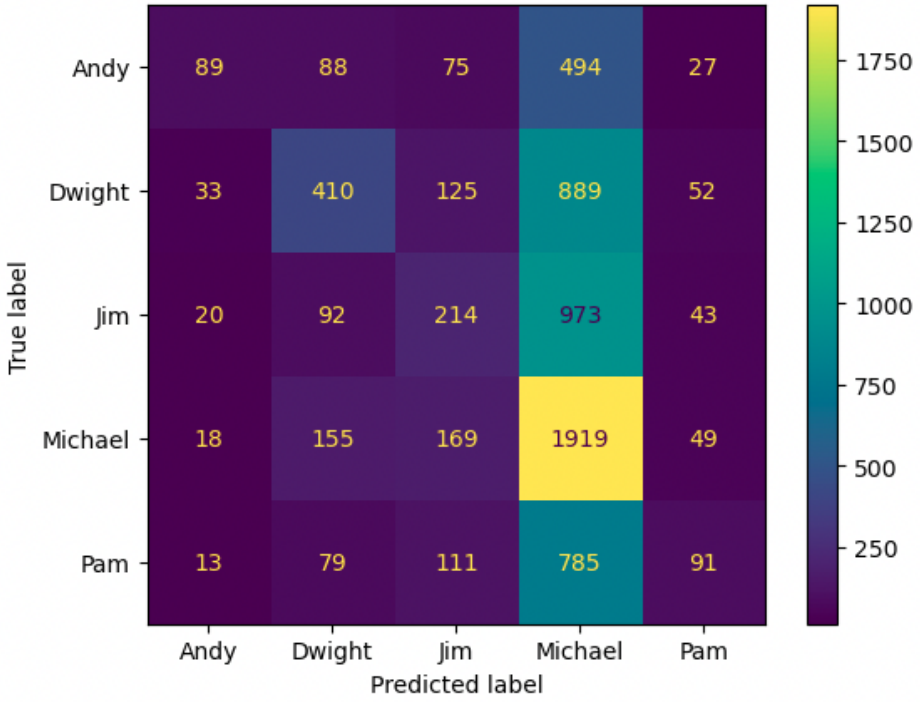
Character	Season	Total Words
Michael	1	12,548
Michael	2	34,503
Michael	3	29,159
Michael	4	22,009
Michael	5	27,617
Michael	6	26,062
Michael	7	20,630
Andy	8	14,339
Dwight	9	14,332

Question 3

ML Algorithm Accuracy

Naive Bayes ML Model Accuracy	38.83%

Confusion Matrix



Question 4

Overall tone of the script

Sentiment Label	Sentiment Score
POSITIVE	0.6571989059448242

Question 5

Top 10 characters with the most lines

Character	Number of Lines
Michael	9207
Dwight	5681
Jim	4601
Pam	3651
Andy	3045
Kevin	1188
Angela	1175
Oscar	1073
Erin	1047
Ryan	977

Bottom 10 characters with the least number of lines

Character	Number of Lines
3Rd Athlead Employee	1
Meredith & Kelly	1
Member	1
Meemaw	1
Mee-Maw	1
Man In Video	1
Man 3	1
Man 1	1
Male Voice	1
Meredith’S Vet	1

Question 6

Top phrases used by main character

Phrase	Number of Times Used
no no no no	112
no no no no no	47
we re going to	39
we are going to	37
okay you know what	35
you know what you	33
what are you doing	32
and you know what	27
the end of the	25
all right all right	25
that what she said	22
thank you very much	22
in the conference room	22
you re going to	21
well you know what	20
good to see you	20

Team contributions

Jenny: Wrote code to answer questions 1, 2, 3. Wrote the method of analysis, program description, and part of the conclusion for these questions. Wrote results for these questions.

Patrick: Wrote code to answer questions 4, 5, 6. Wrote the method of analysis, program description, and part of the conclusion for these questions. Wrote results for these questions. Wrote code for the word cloud.

Conclusion

The Office dataset was analyzed to address multiple questions. The findings revealed interesting insights. Firstly, it was determined that the character "Michael" spoke the most words throughout the entire show. Moreover, this pattern persisted consistently from seasons 1 to 7. However, in seasons 8 and 9, the characters "Andy" and "Dwight" take the lead.

Regarding question 3, a Naive Bayes ML model was employed to classify spoken lines to the respective characters. However, the model's accuracy rate was disappointingly low, reaching only 38.83%. This outcome suggests the algorithm struggled to classify the characters based on their dialogue correctly. The confusion matrix provided a visual representation of the comparison between predicted and actual results. Notably, the algorithm frequently over-predicted for the character "Michael" while under-predicting for the remaining four main characters: "Andy," "Dwight," "Jim," and "Pam."

Another interesting insight was the overall tone of the show. The program characterized the tone of the script overall as positive, with a positivity rating of 65.7% out of 100%. This is to be expected from a TV sitcom such as The Office, but it is more interesting to be able to quantify its positivity.

The analysis also revealed a large number of characters with only one line, while main characters like Michael, Pam, and Jim had several thousand lines across the entirety of the script. While some of the common phrases said by Michael were more mundane, such as “thank you very much” and “good to see you,” a few of the phrases more uniquely characterize the show, such as “in the conference room” and “that’s what she said.”

The dataset clearly establishes "Michael" as the primary and central character. This is evident from the character's substantial word count throughout the show, with thousands of lines, as well as the character's consistent dominance in the first seven seasons. However, due to the consistent prominence of "Michael's" dialogue, the Naive Bayes model developed a bias and tended to overclassify lines under that character.

Reference(s)

Cominetti, F. (n.d.). *The Office Lines*. Kaggle. Retrieved May 29, 2023, from https://www.kaggle.com/datasets/fabriziocominetti/the-office-lines?datasetId=1807639&select=the-office_lines.csv

OfficeQuotes.net. (n.d.). *Welcome to OfficeQuotes.net*. OfficeQuotes.net. https://www.officequotes.net/