Title: Project 1 Milestone 2
Group: Panic at the Deadline
Group Leader: Jessie Bailey
Group Members: Ariana Elahi, Sean Bamfo
Course: DS 4002-001
Date: 9/15/2025

Hypothesis: A machine learning model trained on dialogue data from *South Park* will be able to correctly identify the character who delivered a randomly chosen line with at least ~90% accuracy.

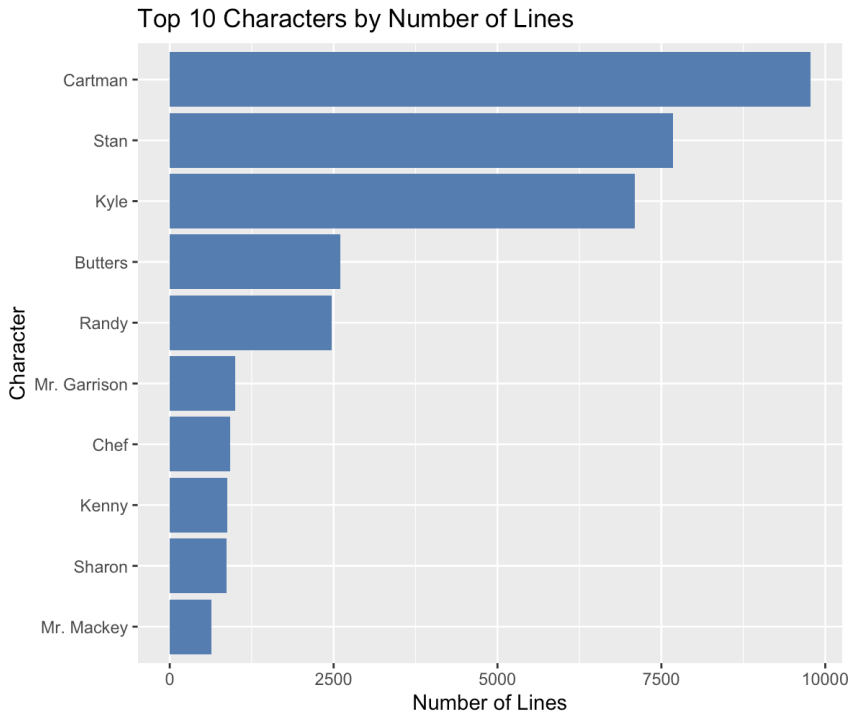
Research Question: For a randomly chosen *South Park* line, can we determine the character?

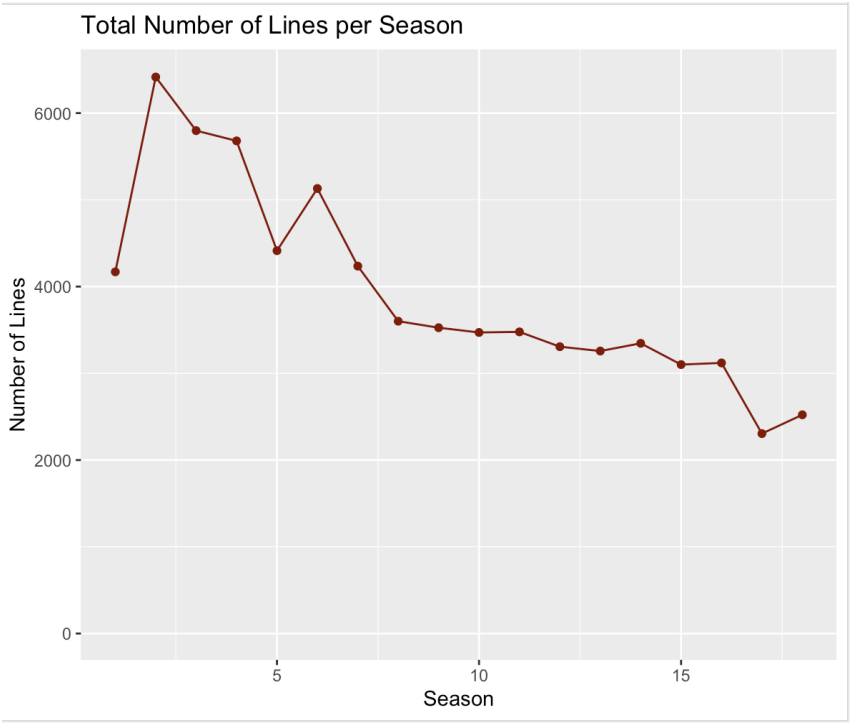
Model Approach: In order to model our goals, we can first take all the *South Park* lines and clean them up so each line is linked to the character who said it. To guess which character might say a new line, we can turn the words into numbers, using methods like basic word count, and train a logistic regression model, which is good at text prediction. To see which characters show up most over time, we can count how many lines each character has in each season and then use a simple line-fitting method, like a different linear regression, to see trends and make guesses about who might appear more often in future seasons. By using this method, we are able to use word count and line counts to represent simple models in this ecosystem that may be able to predict which characters said what lines.

Executive Summary: This document details the dataset establishment and analysis plan for the project defined by the research question above. The dataset establishment section includes exploratory plots, a summary of the data, the data provenance, the data license, the data dictionary, an ethical statement, questions regarding the dataset, and hypothesis refinement. The analysis plan section contains a graphical depiction of the analysis plan, the steps to the plan, and a specific goal to mark the finish line of the project.

Dataset Establishment:

- Exploratory plots





- Summary of data
    - The data is a .CSV file containing information regarding the season, episode number, character, and a given line from *South Park*. Data is text strings, with the exception of the season number and episode number, which are numerical values. The data can be downloaded from the team's shared GitHub repository, linked here:

https://github.com/Seanhru/PanicAtTheDeadline [1]. There are two data files, the original (named All-seasons.csv) and the cleaned data (named filtered_data.csv).

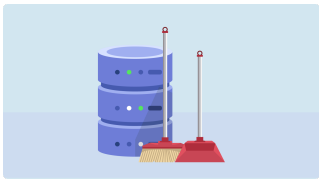
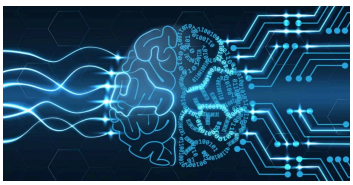
- Provenance
    - The dataset used in this project contains seasons, episodes, characters, and lines from the television show *South Park*, created by Trey Parker and Matt Stone. While the majority of the show's episodes were written by Parker and Stone, additional contributions from developers, directors, and guest writers are listed on the *South Park* page of IMDb [2]. The original dataset was compiled by Bob Adams from SouthPark.Fandom.com, containing data from seasons 1 through 18, and is publicly available on GitHub at: https://github.com/BobAdamsEE/SouthParkData [3] [4]. The data originated as a .HTML file, but was modified by Adams for reformatting to a .CSV file and the correction of typos. On September 12, 2025, the team downloaded the dataset from Adams' repository and transferred it to their own repository (linked in the summary section above). Since that time, the team has made modifications to the dataset. All subsequent changes and related work are documented in the team's repository.
- License
    - The data is licensed under *Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)* in the Creative Commons.
- Data dictionary
    -

Column	Description	Data
Season	The season number of a particular episode of *South Park*. Numerical value.	Any number 1 through 18.
Episode	The episode number of a particular episode of *South Park*. Numerical value.	Any number 1 through 17.
Character	The name of the character who delivers a particular line. Text string.	Example characters: Cartman, Stan, Kyle, Butters, Randy, Mr. Garrison, Chef, Kenny, Sharon, Mr. Mackey.
Line	The line that a particular says during a particular episode. Text string.	Example line: *You guys, you guys! Chef is going away.*

- Ethical statement
    - All *South Park* characters, themes, and dialogue are the property of Trey Parker and Matt Stone. While *South Park* may contain certain commentary or humor that is inappropriate for an educational setting, the team strives to focus instead on the patterns found in the data to make a functional machine learning model.
- One question explored and answered

    - One question we explored was how the number of lines in the script was impacted and changed according to the seasons and other characteristics, like character. Through our data establishment, we were able to find that as the seasons progressed, the number of lines spoken decreased, displaying that there became a heavier emphasis on non-verbal screen time. Additionally, through the line count per character, we were able to establish the 10 ten most important and critical characters to the show such as Cartman, Stan and Kyle, who have significantly higher line numbers than other characters.
- One current question and where it is resolved/mitigated in analysis plan
    - One question that we have not addressed yet is how the content of the script lines has changed over the seasons, and by character. We can resolve this issue by searching for the most common words in the entire series, and searching how the words counts for those specific words change over all seasons.
- Refinement of hypothesis/research question/model approach
    - We would like to refine our hypothesis to make it more attainable. Although having a 90% accuracy would be great, we would like to modify it to having at least a 70% accuracy because in a class setting, above 70% is the standard way of determining pass versus fail. Considering we want to have a model that is consistent with class policy, this modification seems reasonable.
- Goal, put another way
    - Our goal is to create a machine learning model that is able to tell what *South Park* character is most likely to be the speaker, given a specific line. Our original data set can be found here: https://github.com/BobAdamsEE/SouthParkData [4]. Our filtered data set is here: https://github.com/Seanhru/PanicAtTheDeadline [1]. We will be training a machine learning model and uploading it to the latter GitHub link. Any person who has this second github link will be able to recreate this project as all of our code will be there.

Analysis Plan:

- Graphic of analysis plan
    -  →  → 
    - Data Cleaning → Train Model → Test Model
- List steps of plan, each step is 1 paragraph
    - Preprocessing
        - For our first step, the preprocessing stage, the dataset should first be filtered to focus on the top 10 characters ranked by line volume, ensuring that the analysis emphasizes the most prominent figures. It is also important to document key structural details of the dataset, including the total number of rows, the number of distinct episodes represented, and the range of seasons covered, as these provide context for the scope and completeness of the data. A comprehensive data dictionary should then be created to define each variable, clarify data types, and explain any derived fields or transformations applied. Finally, any uncertainties,

gaps, or major issues within the dataset, such as missing values, inconsistent formatting, or potential biases, should be clearly noted to highlight limitations and inform subsequent analysis steps.

- Data Cleaning
  - The data cleaning process involves standardizing the text in the line column by converting all dialogue to lowercase, removing punctuation, and stripping any unnecessary whitespace to ensure consistency. Lines with ambiguous or missing speaker attributions should then be either removed or flagged for review, as these could introduce error into the analysis. Once the text is standardized, dialogue lines can be tokenized and used to extract n-grams, which capture common word patterns and phrases. Finally, the dataset should be split into training, validation, and test sets, with stratification by character to maintain balanced representation across subsets and reduce bias in model evaluation.
- Exploratory Data Analysis
  - The exploratory data analysis (EDA) stage will begin with visualizing the line counts per character, such as through a bar plot, to highlight which characters dominate the dialogue and how representation varies across the dataset. Next, we want to analyze the distribution of line lengths and overall vocabulary usage, which can provide insights into stylistic differences between characters as well as the diversity of language used. It is also important to check for class imbalance, since large disparities in dialogue volume between characters may skew model performance. Finally, potential language or thematic drift across seasons should be examined, both to understand how dialogue evolves over time and to anticipate challenges for the train, validation, and test split, where shifts in tone, vocabulary, or character focus could impact generalizability.
- Modeling
  - The modeling stage will start with training a multiclass logistic regression model using TF-IDF features, as this serves as a strong baseline for text-based classification tasks. Additional models such as linear regression and the multinomial Naive Bayes classifier, can then be tested to provide alternative benchmarks and insights into how different algorithms handle the dataset. To further enhance performance, more advanced methods like random forests, support vector machines (SVMs), or lightweight neural architectures may be explored, as they can capture nonlinear relationships and more subtle text patterns. All models should be evaluated not only on overall accuracy but also on per-class precision, recall, and F1 scores to account for potential imbalances across characters. A confusion matrix should also be generated to visualize misclassifications and better understand where models struggle, guiding refinements in preprocessing or feature engineering.
- Evaluation
  - For evaluation, examining which words, phrases, or n-grams have the strongest influence on predictions makes it possible to identify the linguistic cues most associated with specific characters. This not only provides interpretability, which reveals what drives character attribution, but also offers insights into

character-specific language patterns and thematic tendencies within the dataset. Such analysis can highlight whether the model relies on meaningful features or is biased toward superficial patterns, helping to validate both the robustness and fairness of the results.

- Specific, quantifiable goal
    - We will know that we hit our goal and have passed our project finish line when the model is given a specific line and can predict what character most likely said it with at least 70% accuracy.
- Goal, put another way
    - This project aims to build a machine learning model that predicts which of the top 10 *South Park* characters spoke a given line, achieving at least 70% accuracy through systematic preprocessing, exploratory analysis, and model evaluation.

References:

[1] Seanhru, "GitHub - Seanhru/PanicAtTheDeadline: DS4002 Project 1," GitHub, 2025. https://github.com/Seanhru/PanicAtTheDeadline (accessed Sep. 15, 2025).

[2] "South Park (TV Series 1997–) - Full cast & crew - IMDb," IMDb, 2025. https://www.imdb.com/title/tt0121955/fullcredits/ (accessed Sep. 15, 2025).

[3] "South Park Public Library," Fandom.com, 2025. https://southpark.fandom.com/wiki/South_Park_Public_Library

[4] BobAdamsEE, "GitHub - BobAdamsEE/SouthParkData: .csv files containing script information including: season, episode, character, & line.," GitHub, 2016. https://github.com/BobAdamsEE/SouthParkData (accessed Sep. 15, 2025).