Title: Project 1 Milestone 1
Group: Panic at the Deadline
Group Leader: Jessie Bailey
Group Members: Ariana Elahi, Sean Bamfo
Course: DS 4002-001
Date: 9/12/25

Motivation: Many TV shows only last a few seasons; however, *South Park* has managed to remain on air for 28 years, winning multiple Emmy Awards despite its characteristically controversial comedy [1]. Its longevity is especially notable because the show has continuously adapted its humor, social commentary, and characters to remain culturally relevant across generations. Consistent and likeable character identity and voice are crucial aspects of any story, enabling the audience to connect with the plot on a more personal level [2]. In *South Park*, each character has a highly distinctive speaking style, vocabulary, and set of recurring themes, making dialogue one of the most recognizable aspects of the show. Character-specific language is not just a comedic tool but a critical element of the show's cultural endurance.

Utilizing data science and machine learning, this project seeks to analyze the type of dialogue associated with each character in *South Park* to determine what characteristics, mannerisms, and speech patterns resonate with viewers to keep them engaged and entertained. Additionally, the project can identify recurring patterns in humor and storytelling techniques across seasons and draw links between season success and effective script writing.

Goal Statement: We will build a model that can identify the *South Park* character based on a randomly chosen line by using machine learning, which will be able to correctly identify the character 90% of the time.

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

Modeling 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.

References:

[1] The Editors of Encyclopaedia Britannica, "South Park | Characters & Description | Britannica," Encyclopædia Britannica. 2019. Available: https://www.britannica.com/topic/South-Park-television-series

[2] S. Hardin, "Who Goes There? The Importance of Writing Distinct Character Voices," Medium, Jun. 2022. https://medium.com/ironsource-levelup/who-goes-there-the-importance-of-writing-distinct-character-voices-2d5260d05e56

Data: https://github.com/BobAdamsEE/SouthParkData/blob/master/All-seasons.csv