Final Project Proposal
- South Park Character Poem Generator

COSC-572
Empirical Methods in Natural Language Processing

Group Member
Steven Chang (yc704), Tianrun Sun(ts1158), Yushi Zhao (yz521)

Abstract
South Park owns the most interesting dialogues among all cartoons. We all know that South Park is a cartoon with strong character personalities. By using the dialogue dataset, we will be able to build a generator that teaches the South Park characters to write poems in their own language or speaking style. At the end of the training session, they will all be able to write poems that reflect their mood and how they feel in their personal style.

Dataset
Data download from kaggle
Size: 5.28MB
Details: 143672 lines, annotated with season, episode and speaker
Link: https://www.kaggle.com/tovarischsukhov/southparklines/data

Application:
Firstly, we try to build an n-gram model for each character, and generate sentences with that model. These sentences are then filtered by the line-final rhyme that the user defines. Finally, we put sentences together to generate poems.
Furthermore, we plan to analyze the sentiment of each generated sentence. It is implemented by choosing “Bing”, "nrc" and "afinn" lexicons to get the sentiment score for the word token in the sentences for each character. Why we chose three different lexicons? Because each of the lexicons can be complex or simple, ex. “Bing” only categorizes into positive and negative while “afinn” matches the word to the number between -5 to 5.

Work Division
Steven: Sentiment Score Analysis
Yushi: Rhyme Analysis and Poem Generator
Tianrun: Model Generator

Timeline
4/2 - 4/8: Understanding and Clean data. Decide and define the model.
4/9 - 4/13: Build model and generator(n-gram).
4/16 - 4/20: Poem generator with sentiment score.