Math 189R Project Proposal

Shanni Lam and Marcelo Almora Rios

June 15, 2020

Using a poetry dataset of 3,000,000 poetry lines, we will extrapolate the words from each poem. This will be accomplished with Python libraries such as pandas, matplotlib, numpy, scipy, textblob, word2vec, and more. We will explore ideas like whether certain combinations of words are more poetic than others and recognition of poems by emotion, shape, meter, and rhyme. We may also try to create edited implementations of machine poetry generation from these features. First, we will read the current literature about such natural language processing topics. Some interesting papers are linked below. Afterwards, we will use big data machine learning techniques, including semantic analysis, principal component analysis, topic modelling, latent Dirichlet allocation, and neural networks to create a model so that the machine generates its own poem based off the millions of poems in the training set.

The dataset in question is linked here: https://github.com/aparrish/gutenberg-poetry-corpus

The papers in question are linked here:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.730.9340rep=rep1type=pdf