CS 6140 Project Proposal

Jim Morgan, Foad Namjoo, Ariel Oconitrillo

Introduction

"What's in a name? That which we call a rose by any other name would smell just as sweet." These words, penned by William Shakespeare, convey how little Juliet's family name means to Romeo, who would overlook their family feuding for romance's sake. The quote can also apply to the use of language in Shakespeare's plays. Foad, Jim, and Ariel have elected to dive deeper into the works of William Shakespeare and perform various data mining techniques on them.

Data Provenance

We plan to use a dataset comprised of lines of dialogue and stage directions in Shakespeare's plays. The dataset can be found on Kaggle's website here.

Data Structure

From an initial position, we plan on performing an n-gram analysis of Shakespeare's various plays, looking for similarities in word usage and sentence structure between plays. Of additional interest are the influences of Shakespeare's works on later plays and films. It would be interesting to perform a page-rank analysis of Shakespeare's work with modern media like Star Wars. Additional potential targets for comparison include Moby Dick by Herman Melville and A Christmas Carol by Charles Dickens.

Problem Statement

The works of William Shakespeare are of profound importance to all English speakers and are part of the core curriculum in English-speaking countries. Shakespeare's word usage, construction of the plot, and usage of performative techniques like the soliloquy to convey a character's mental and emotional state were all novel devices at the time. However, some questions remain about Shakespeare's authorship, such as whether they were penned by one man or several. It is hoped that through n-gram analysis, we can better qualify Shakespeare's works and look for consistencies between plays.

Learning Objectives

While it is likely little room for improving the professor's understanding of the data mining techniques used, it is hoped that there will be room for learning more of Shakespeare's literary techniques and how his word choice and sentence structure influences modern language to the current day.