**Project Type:** Data Analysis

**Project Description:**

This project will be an analysis of data scraped from the fanfiction site “fanfiction.net”. Briefly, fanfiction is a fan-written story about a piece of media--the story can be anywhere from 100 words to 100,000 words, usually taking the characters of the world and writing scenarios that are not present, or would otherwise be impossible, in the original work. In this project, I will be focusing on the metadata of the site’s archive, finding patterns that may arise given the relationship between the genre/category of a work, its word count, and how much engagement it gets (measured in amount of reviews, “favs”, and follows). The primary purpose of this project is to see if there is a correlation between a work’s category, word count, and level of engagement, and if so, what that correlation looks like.

**Proposed Methodology:**

The problem I have outlined requires a methodology that illustrates relationships rather than classifies, so I will be using clustering to measure: 1) if certain categories of works receive higher levels of engagement; and 2) if a higher word count indicates a higher level of engagement.

**Techniques:** I will be using k-means clustering to identify the patterns.

**Approaches:**  
 This project will be framed using an exploratory approach rather than a hypothesis-driven one. By using a more exploratory approach, I hope to gain wider insights into the data.

**Implementation Choices:**

Assuming this refers to “how I will be using the results of the analysis,” if there is a correlation between a work’s category or word count and its level of engagement, it would theoretically solve a problem that some fanfiction writers have stated where a lack of engagement leads to a drop in desire to continue writing. If “implementation choices” refers to where I will be placing the analysis and algorithms, I will be using Google Colabs for the majority of the analysis in a Python notebook due to a prior history of Jupyter notebook hiccups.

**Resources Used:**

<https://www.kaggle.com/metrovirus/fanfictionnet> for the dataset

Excel to clean the data

Google Colabs (Python, Pandas, SciKit) to create the algorithm(s) and analyze the data

**Tentative Schedule:**

W1 (10/20-10/26) - Cleaning the data, begin preprocessing

W2 (10/27-11/2) - Preprocessing continues, lead into reduction and projection

W3 (11/3-11/9) - Algorithm writing, debugging, testing

W4 (11/10-11/16) - Algorithm writing, debugging, testing / start data mining

W5 (11/17-11/23) - Data mining

W6 (11/24-11/30) - Data mining and interpretation

W7 (12/1-12/7) - Data mining and interpretation / consolidation of knowledge

W8 (12/8-12/14) - Anticipated completion this week or the week prior, finalize report

**Data Description and Sample:**

The dataset contains information scraped from the fan fiction archival site “fanfiction.net”. Features include the original media type, the original media’s title, the name of the fan fiction’s author, the location of the work on the site, the story ID, the user ID, the language of the work, the category (or genre), its rating, the amount of chapters, the review count, the word count, the “fav” count, the “follow” count (these are people who request to be notified when the story updates), the date published, the date updated, and the description of the work as written by the author. Some essential data is missing, as can be seen in the data sample below--some works do not have a category, and others are missing dates. The dates are in a format that is impossible for Excel to read (they are listed on the website as, for example, “Oct 19, 2019” or simply “Oct 19”). There are a total of 8,314 separate csv files in the dataset, and each file is a container holding the fan works created for one piece of media.

See below for a screenshotted sample of the data.

