**Data Science Capstone 2 – Project Proposal**

**Problem Statement**

The aim of this project is to develop a recommendation engine that outputs five movies, based on the input of a single movie. I will be using clustering, natural language processing, and similarity measures in order to construct my recommendation engine, and will be using data from the MovieLens dataset and from IMDb to test its efficacy.

**Data Sources**

1. <https://www.kaggle.com/rounakbanik/the-movies-dataset>
2. <https://www.imdb.com/>

I will be combining information from both of these datasets to produce the necessary insights. This will require considerable data manipulation in Python using the pandas module.

**Context**

Movie streaming is currently an extremely popular way to access movies. When someone browses a movie on the web, this is an opportunity to direct them to movies that they will potentially enjoy based on the movie currently being browsed. The site on which the movie is being browsed could benefit from an accurate recommendation system in order to place targeted ads for its users.

**Criteria for Success**

My primary criterion of success will be developing a recommendation engine that provides movies that people would be expected to want to watch, based on the available data for testing. I will be using the MovieLens ratings dataset to this end. The higher (after standardization) the average rating for the movies recommended, the more performant the system. Other metrics such as the likelihood of whether the user has already seen the movie will also be considered. Deliverables for this project will be a report containing my findings, including graphics, and a working recommendation engine that can be applied to any movie that is part of the subset of the MovieLens dataset that I will be working on, and that can be extended to any movie added to the database.

**Scope of Solution Space**

The focus will be a subset of the movies found in the MovieLens dataset. Specifically, I will be limiting the ratings set to users who have rated 1000 movies or more, and movies with at least 50 ratings. This will allow for a robust testing environment, while retaining the majority of movies that people are likely to be interested in.

**Constraints**

The MovieLens dataset only contains movies up to 2018, so testing will be limited to them. However, the recommendation engine itself will be using IMDb data to provide the recommendations, so will be able to remain current.

**Stakeholders**

The owners of the websites who would be employing the recommendation engine.

**Outline of Solution**

Initially, I will analyze the MovieLens dataset, and acquire the data I need from IMDb via web scraping in Python. I will then perform EDA on the data available to gain initial insights. I will use hierarchical clustering to establish relationships between movie genres and sub-genres, and divide the movies into these categories. I will perform a similar step with the decade in which the movie is released. In order to have the capability to keep the recommendation engine current, it can’t rely on MovieLens data, which is not consistently updated, so I will scrape reviews from IMDb for each movie, and determine similarity via keywords found in these reviews. I will finally construct my recommendation engine, incorporating the above analysis into the code. I will test the system on the movies found in the MovieLens dataset, and record the results.