**Design Document**

**Goals:**

**Long Term Goal:** To successfully create a model to accurately generate movie recommendations from movies released before July 2017 based on certain features

**Intermediate Goal:**

* Create data frame for each csv fie and merge them
* Explore the data to find any missing information
* Clean up data
* Create features

**Process:**

* Data Acquisition:
  + Download data
  + Turn dataset into data frame
* Data Preparation:
  + Display data
  + Merge data
* EDA (Exploration of Data analysis):
  + Find any missing data
  + Expunge any data that might mess with the dataset
* Model development:
  + Define a function to compute the scores
  + Filtering the movies to be scored
  + Compute the score for every movie in the dataset
  + Define a TF-IDF Vectorizer Object. Remove all english stop words.
  + Construct the required TF-IDF matrix by fitting and transforming the data
  + Compute the cosine similarity matrix
  + Define Function that takes in movie title as input and outputs most similar movies

**Project Roles:**

Program leader: Samuel

Schedule to start: Throughout the Capstone

Lead Code Designer: Jermaine

Schedule to start: Throughout code Development

Note taker: Nasir

Schedule to start: Throughout the Capstone

Auditor: Samuel

Schedule to start: Once the code has been completed

Presentation Designer: Antoine

Schedule to start: Throughout the capstone

**Problem space:**

* Data was now as cleaned as we needed there are rows with numerous missing information which messed with our data.
* Ran out of ram when running full code
* Code could not run in VSC
* Datafile was too large to run in VSC