Recommender Systems-- Final Project Proposal

Violeta Stoyanova, Peter Kowalchuk

**Introduction:**

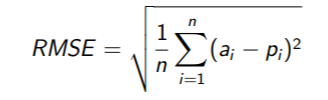
“For those of us fortunate enough to be healthy and working from home during the outbreak of COVID-19, the possibilities for entertainment can feel both extremely confined (specifically, to our homes) and overwhelmingly endless (where in our Netflix queues to begin?) (Cohen, 2020).” Now more than ever in our socially-distanced life we depend on recommendations for movies to fill our quarantine days with entertainment. Some rely on movie experts; others look at ratings to pick flicks and cure some of the overwhelming boredom. In this regard, for our final project we will attempt to build a recommender system with the use of ‘implicit’ movie data for the model to recommend new movies to users.

**Objective:**The goal of this project is to build various models in the attempt to recommend new movies to users based on ‘implicit’ data. We will be using a dataset from kaggle.com that contains descriptions of 34,886 movies from around the world. This dataset contains a summary of the plot of each movie. The goal is to use this plot to generate recommendations based on the works present in the plot. A first implementation of the recommender will provide a recommendation based on a word or list of words provided. One can imagine how a more complex implementation of this approach could allow the user write a plot of a movie she/he would like to watch. The system then would provide recommendations based on this plot by using the list of words contained in it as input to the recommender system presented in this project.

**Methodology:**

For the purposes of this project we will utilize the speed and scalability of Apache Spark via Databricks. The first step will be to tokenize the data, so we are able extract some weight with the use of the ‘genre’ variable. Via TF-IDF we will first eliminate any stop words and count word frequency and the build a utility matrix. With this approach words will be equivalent to users in a conventional recommender, with the word weight provided by TF-IDF representing the ratings for each movie. Afterwards, using the techniques we have learned and applied in class, such as Item-based collaborative filtering and User-based collaborative filtering (IBCF and UBCF, respectively) and ALS (Alternating Least Squares) we will build recommender systems in order for movies to be recommended to users.

We believe that proper evaluation is important to detect how accurate our algorithms are at recommending movies thus we will root mean square error metric (RMSE).



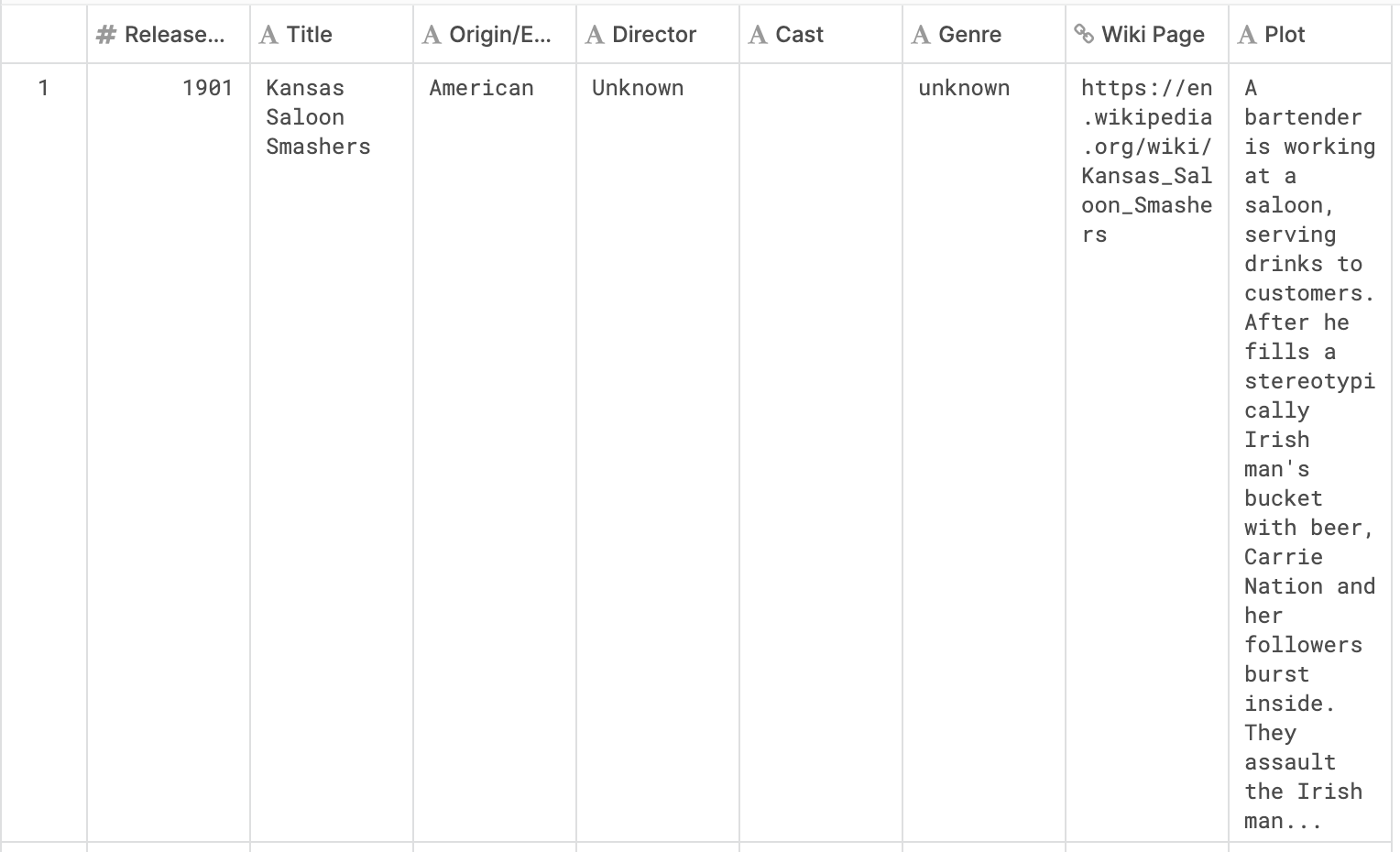
Where the lowest RMSE will indicate the better model.

**Data Description:**

Content

The dataset contains descriptions of 34,886 movies from around the world. Column descriptions are listed below:

* *Release Year* - Year in which the movie was released
* *Title* - Movie title
* *Origin/Ethnicity* - Origin of movie (i.e. American, Bollywood, Tamil, etc.)
* *Director* - Director(s)
* *Plot* - Main actor and actresses
* *Genre* - Movie Genre(s)
* *Wiki Page* - URL of the Wikipedia page from which the plot description was scraped
* *Plot* - Long form description of movie plot (WARNING: May contain spoilers!!!)



**Data Sources:**

<https://www.kaggle.com/jrobischon/wikipedia-movie-plots>