# Research Project Proposal

Your Name
Date

## Machine Learning in Movies of IMDB

I'd like to apply machine learning algorithms such like classification or clustering to movies, which will be collected from the IMDB. I will try to scrape the data from IMDB, store and clean it. Then try to find some interesting points from the dataset and apply machine learning algorithms to it. So far, I am sure I will apply classification and regression to it.

## Background

I am a crazy movie fan, so I want to imply my machine learning skills to categorize them and to find interesting things among them.

## Date sources

**http://www.imdb.com**

### Algorithm:

1.Linear Regression

In this post I will implement the linear regression and get to see it work on data. Linear Regression is the oldest and most widely used predictive model in the field of machine learning. The goal is to minimize the sum of the squared errros to fit a straight line to a set of data points.

2.Logistic Regression

In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical. This article covers the case of binary dependent variables—that is, where it can take only two values, such as pass/fail or win/lose. Cases with more than two categories are referred to as multinomial logistic regression, or, if the multiple categories are ordered, as ordinal logistic regression.

Logistic regression was developed by statistician David Cox in 1958(although much work was done in the single independent variable case almost two decades earlier). The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). As such it is not a classification method. It

could be called a qualitative response/discrete choice model in the terminology of economics.

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it treats the same set of problems as probit regression using similar techniques, with the latter using a cumulative normal distribution curve instead. Equivalently, in the latent variable interpretations of these two methods, the first assumes a standard logistic distribution of errors and the second a standard normal distribution of errors.

**References**


[1] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: An experiment in text regression. In Proceedings of NAACL-HLT, 2010.

[2] Leonid Velikovic, et al. “The viability of webderived polarity lexicons.” NAACL, 2010.

[3] https://en.wikipedia.org/wiki/Logistic_regression