FILM SCRIPT ANALYZER, BY LUIS CASTRO

Rationale:

Mathematics provides a framework to understand the concepts and parameters by which a work of art is created. Literature, music, paint and sculptures can be understood through statistics, by finding underlying patterns and properties that are intuitively transmitted, but are difficult to grasp independently.

Film industry, the cinema, the seventh art brings many facets of artistic production together, and is able to create marvels that overwhelm our senses. It is also a billion dollars industry that has great impact on society and culture.

The root, the germ seed of a movie is the script. Scripts are the recipes, the framework for the director to bring his/her vision to life. Professionals and amateurs alike write thousands of scripts, many have the potential to be gold, and others are worthy of the trash bin.

Scripts are read and can be time consuming to be read in their entirety, and even after that and after trained and skillful directors and producer have their opinions of them, many movies fail in the eye of the public and thus, in their box office returns.

There should some way to speed up the process and have a guide to choose the best scripts, to have a certain degree of confidence that a movie will be well received.

Methodology:

Natural Language Processing is a powerful tool. With the increasing processing power available, the fast rate at which structured and non-structured data is produced, and the interconnectivity and availability of data from various sources.

"The most merciful thing in the world, I think, is the inability of the human mind to correlate all its contents." H.P. Lovecraft

Steps:

1) Scrape the web for available film scripts.
2) Obtain meta-data of films using IMDB's API, here is found also our target values. The IMDB rating that is a democratic, worldwide poll of movies. It can show the impact and acclaim of a movie.
3) Clean and format the information to be usable.
4) User IBM Watson's API to perform a Personality Insight of the scripts. It assigns texts with a set of 30 numerical features describing them.
5) Obtain additional features using NLP (by means of the NLTK package). This package provides tools to perform numerical, semantic and syntactic analysis of text.
6) Perform exploratory analysis of data, uncovering most relevant features.
7) Use supervised learning algorithms (linear regression, support vector machines, k-nearest neighbors, neural networks, etc.) or an ensemble learner to have cross-validated predictions of the rating.
8) Keep increasing the size of the dataset and tuning parameters.
9) The result will be a robust model that can accurately predict the IMDB rating of a script, can provide a baseline for user (say only read scripts with > 70 predicted rating).
10) An added result can be the interactions with other variables in film production like best actor/actress or Director for a film.