IMDB Database

Six Degrees of Kevin Bacon

Basic Idea of Project

- Download and parse through the IMDB movie databases to extract data for all living actors / actresses.
- Using Databricks / Spark, compute various queries such as: the number of films for each actor, Pagerank of each actor (similar to how Google ranks relevance of webpages), and find degrees of separation of each actor between them and actor “Kevin Bacon”.

Machine Learning Aspect

- After computing some basic queries, the final task was to build and tune a linear model to predict the IMDB rating of any given movie based on the “quality” of the cast members. These quality values are the same as the queries described earlier, such as number of movies appeared in, Pagerank, and distance from Kevin Bacon.

Skills Learned

- The entire Databricks workflow, from downloading raw data, to combining it into a single data table, to cleaning it, to performing search queries, to applying machine learning algorithms to solve predictive problems.
- Connecting to Databricks cloud service and working and saving remotely.
- SQL / Spark language skills