**Termahter Picker**

Mohamed Sleem, Luis Perozo

In the scope of our project, the main source of data for our algorithm comes from IMDB json queries for each movie title already released and rated. So ideally, the range of data that we can acquire is a json file with all the movies from 1950 to 2012 which can be used as training data.

As of this moment, we have pulled all the movie titles from the year 1950 to 2012, and have ran scripts to produce json files from the IMDB api mentioned in our proposal. Naturally, some of the queries would not match due to name conflicts. However, we were successfully able to pull about 11,292 movies and their information into our data collection. The total number of actors in these movies totaled to almost 58,000 unique actors/actresses. Here are some further statistics regarding our data:

**Example data format, Batman: The Dark Knight:**

{"rating": 9.0, "genres": ["Action", "Crime", "Drama", "Thriller"], "rated": "PG\_13", "language": ["English", "Mandarin"], "title": "The Dark Knight", "poster": "http://ia.media-imdb.com/images/M/MV5BMTMxNTMwODM0NF5BMl5BanBnXkFtZTcwODAyMTk2Mw@@.\_V1.\_SY317\_CR0,0,214,317\_.jpg", "imdb\_url": "http://www.imdb.com/title/tt0468569/", "directors": ["Christopher Nolan"], "also\_known\_as": ["Batman: The Dark Knight"], "imdb\_id": "tt0468569", "country": ["USA", "UK"], "filming\_locations": "2 International Finance Centre, Central, Hong Kong, China", "writers": ["Jonathan Nolan", "Christopher Nolan"], "actors": ["Christian Bale", "Heath Ledger", "Aaron Eckhart", "Michael Caine", "Maggie Gyllenhaal", "Gary Oldman", "Morgan Freeman", "Monique Gabriela Curnen", "Ron Dean", "Cillian Murphy", "Chin Han", "Nestor Carbonell", "Eric Roberts", "Ritchie Coster", "Anthony Michael Hall"], "plot\_simple": "When Batman, Gordon and Harvey Dent launch an assault on the mob, they let the clown out of the box, the Joker, bent on turning Gotham on itself and bringing any heroes down to his level.", "year": 2008, "runtime": ["152 min"], "type": "M", "release\_date": 20080718, "rating\_count": 832786}

**Total movie count (so far)**: 11,292

**Total actor count in our data:** 57,679

**Most frequent actors:** Robert De Niro - 88 Movies. Christopher Walken - 88 Movies.

**Year with most movies**: 2002

Figure Movie Count per Year

At this point in our project we feel comfortable with the data we have pulled. For future data manipulation we plan on further analyzing each actor and their respective importance in each movie, while cross referencing this with the overall success of the movie. If time permits we also plan to take into account the plot, producers, and other characteristics such as genres into our predictor algorithm.