# **CS839 Project Stage 2 Report**

Pan Wu, Dongqiangzi Ye, Sek Cheong

**Data Source**

We extracted movie description and review information from one on line movie review site and movie retail:

* https://www.imdb.com
* <https://www.walmart.com>

The IMDB contain information about movie reviews while the Walmart contains movie DVDs for sale. Both web sites contain information Such as movie title, synopsis, genre, duration, actors/actresses, etc.

**Entity**

We collected movie information from IMDB and Walmart and obtained the following information:

Schema:

movies\_IMDb(name, duration, genre, release\_date, directors, stars)

movies\_walmart(name, duration, genre, release\_date, directors, stars)

Since both websites essential provide a lot of same movie attributes, we were able to come up with the common schema for both tables. The movies\_IMDb table contains 3043 tuples and the movies\_walmart table contains 3140 tuples.

**Open Source Tools Used**

We use the BeautifulSoup python package. BeautifulSoup allows you to pull data out of HTML and XML files. It provides simple methods for navigating, searching, and modifying the parse tree.