Movie Reviews and Revenues: An Experiment in Text Regression

Mahesh Joshi, Dipanjan Das, Kevin Gimpel and Noah A. Smith
Language Technologies Institute, Carnegie Mellon University
{maheshj, dipanjan, kgimpel, nasmith}@cs.cmu.edu

I. The Story in Short
- Use metadata and critics' reviews to predict opening weekend revenues of movies
- Feature analysis shows what aspects of reviews predict box office success

II. Data
- 1718 Movies, released 2005-2009
- Metadata (genre, rating, running time, actors, director, etc.): www.metacritic.com
- Opening weekend revenues and number of opening screens: www.the-numbers.com

III. Model
- Linear regression with the elastic net (Zou and Hastie, 2005)
  \[ \theta = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^n \left( y_i - (\beta_0 + x_i^T \beta) \right)^2 + \lambda P(\beta) \]
  \[ P(\beta) = \sum_{j=1}^p \left\{ (\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha |\beta_j|) \right\} \]

IV. Features
- Lexical n-grams (1,2,3)
- Part-of-speech n-grams (1,2,3)
- Dependency relations (nsubj,advmod,...)
- U.S. origin, running time, budget (log), # of opening screens, genre, MPAA rating, holiday release (summer, Christmas, Memorial day,...), star power (Oscar winners, high-grossing actors)

Meta
- documentary, running time N, philosophical, boganmen, this series, unlitile exists, testosterone, anticipation
- Will Smith, poop, boogeyman

V. What May Have Brought You Here
- Text features can substitute for and improve upon metadata

VI. Results
- Linear regression with the elastic net (Zou and Hastie, 2005)
- Feature analysis shows what aspects of reviews predict box office success

VII. More Cool Features
- www.ark.cs.cmu.edu/movie$-data
- www.the-numbers.com
- www.ark.cs.cmu.edu/movies-data