Textual analysis of movie popularity

Ioanna Sanida, Gian Luigi Chiesa and Sarah Hiller

Problem

- Predict movie success
- Existing papers already achieve high accuracies
- But previous accounts focus on metadata and machine learning techniques, or texts **about** a movie, often coupled with sentiment analysis
- What we do: use only text available **from** the movie
- That is, the dialogues and a neutral plot summary
- Success is measured via IMDB scores

Problem

- Predict movie success
- Existing papers already achieve high accuracies
- But previous accounts focus on metadata and machine learning techniques, or texts **about** a movie, often coupled with sentiment analysis
- What we do: use only text available **from** the movie
- That is, the dialogues and a neutral plot summary
- Success is measured via IMDB scores

Problem

- Predict movie success
- Existing papers already achieve high accuracies
- But previous accounts focus on metadata and machine learning techniques, or texts **about** a movie, often coupled with sentiment analysis
- What we do: use only text available **from** the movie
- That is, the dialogues and a neutral plot summary
- Success is measured via IMDB scores

Problem

- Predict movie success
- Existing papers already achieve high accuracies
- But previous accounts focus on metadata and machine learning techniques, or texts **about** a movie, often coupled with sentiment analysis
- What we do: use only text available **from** the movie
- That is, the dialogues and a neutral plot summary
- Success is measured via IMDB scores

Problem

- Predict movie success
- Existing papers already achieve high accuracies
- But previous accounts focus on metadata and machine learning techniques, or texts **about** a movie, often coupled with sentiment analysis
- What we do: use only text available **from** the movie
- That is, the dialogues and a neutral plot summary
- Success is measured via IMDB scores

Problem

- Predict movie success
- Existing papers already achieve high accuracies
- But previous accounts focus on metadata and machine learning techniques, or texts **about** a movie, often coupled with sentiment analysis
- What we do: use only text available **from** the movie
- That is, the dialogues and a neutral plot summary
- Success is measured via IMDB scores

Problem

- Predict movie success
- Existing papers already achieve high accuracies
- But previous accounts focus on metadata and machine learning techniques, or texts **about** a movie, often coupled with sentiment analysis
- What we do: use only text available **from** the movie
- That is, the dialogues and a neutral plot summary
- Success is measured via IMDB scores

Background assumptions

Assumptions

- Bag of words
- Pairwise independence

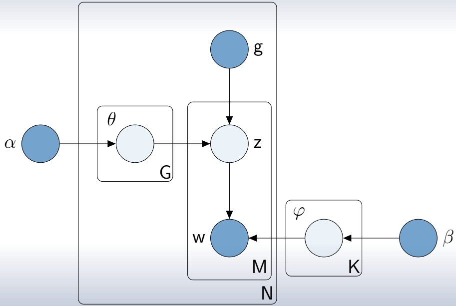
Background assumptions

Assumptions

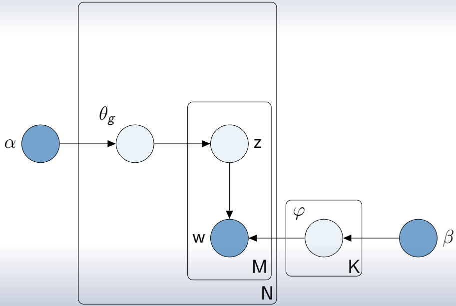
- Bag of words
- Pairwise independence

Generative Model

Generative Model



Submodels (fixed genre)



Dataset

- Combined from two datasets which are available online
- Scripts from Mizil and Lee (2011)
- Summaries from Bamman, O'Connor and Smith (2013)
- Both datasets contain large amount of metadata, including IMDB score

Dataset

- Combined from two datasets which are available online
- Scripts from Mizil and Lee (2011)
- Summaries from Bamman, O'Connor and Smith (2013)
- Both datasets contain large amount of metadata, including IMDB score

Dataset

- Combined from two datasets which are available online
- Scripts from Mizil and Lee (2011)
- Summaries from Bamman, O'Connor and Smith (2013)
- Both datasets contain large amount of metadata, including IMDB score

Dataset

- Combined from two datasets which are available online
- Scripts from Mizil and Lee (2011)
- Summaries from Bamman, O'Connor and Smith (2013)
- Both datasets contain large amount of metadata, including IMDB score

Dataset

- Combined from two datasets which are available online
- Scripts from Mizil and Lee (2011)
- Summaries from Bamman, O'Connor and Smith (2013)
- Both datasets contain large amount of metadata, including IMDB score

Outcome - LDA on “Drama” movies

- n^o movies: 290
- vocabulary size: 35667
- n^o of words: 723036
- n^o of languages: 20
- n^o of iterations: 1000

Outcome - LDA on “Drama” movies

- n^o movies: 290
- vocabulary size: 35667
- n^o of words: 723036
- n^o of languages: 20
- n^o of iterations: 1000

Induced Languages - examples

Language 0: don know ll like just want think ve going did right got tell good time come say didn let

Language 1: president war mr people ve country sir senator general george bob washington kane jim uh chauncey american state army

Language 2: ain ya got gonna don just em goin like nothin doin good man right somethin ma gotta yah yeah

[. . .]

Language 17: harry film movie baxter frances andy marge fran boat kubelik mantan sheldrake christmas eddie white boone da dat famous

Language 18: fuck fucking shit fuckin yeah man gonna money gotta fucked ass guys wanna bitch shut mon cause linda bring

Language 19: alex white truman house jane nathan lila gold al mitchell chief faith dennis epps castor jenny haldeman puff ranch

Induced Languages - examples

Language 0: don know ll like just want think ve going did right got tell good time come say didn let

Language 1: president war mr people ve country sir senator general george bob washington kane jim uh chauncey american state army

Language 2: ain ya got gonna don just em goin like nothin doin good man right somethin ma gotta yah yeah

[. . .]

Language 17: harry film movie baxter frances andy marge fran boat kubelik mantan sheldrake christmas eddie white boone da dat famous

Language 18: fuck fucking shit fuckin yeah man gonna money gotta fucked ass guys wanna bitch shut mon cause linda bring

Language 19: alex white truman house jane nathan lila gold al mitchell chief faith dennis epps castor jenny haldeman puff ranch

Approach 1: Linear Regression

- Training set: Movies 0 – 200, out of 290
- Model: IMDB score $= \beta_1\theta_1 + \beta_2\theta_2 + \cdots + \beta_{20}\theta_{20}$, $\theta_i =$ probability of language i in the movie.

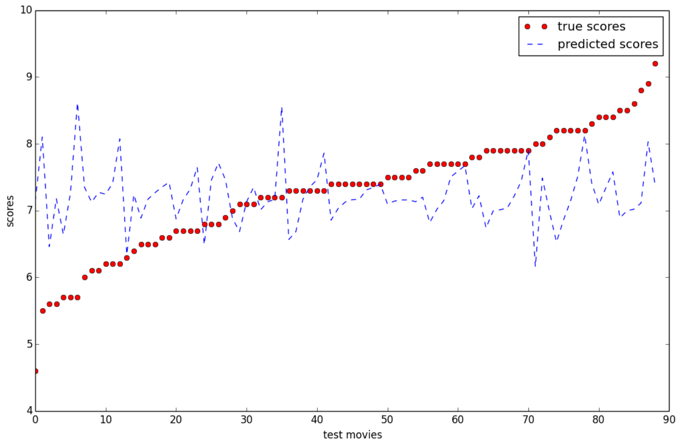
Approach 1: Linear Regression

- Training set: Movies 0 – 200, out of 290
- Model: IMDB score $= \beta_1\theta_1 + \beta_2\theta_2 + \cdots + \beta_{20}\theta_{20}$, $\theta_i =$ probability of language i in the movie.

Approach 1: Linear Regression

- Training set: Movies $0 - 200$, out of 290
- Model: IMDB score $= \beta_1\theta_1 + \beta_2\theta_2 + \cdots + \beta_{20}\theta_{20}$, $\theta_i =$ probability of language i in the movie.

Predicted vs. actual rating, Linear Regression



Approach 2: Logistic Regression

- IMDB score $> 7.4 \mapsto$ successful movie $\mapsto 1$
- IMDB score $< 7.4 \mapsto$ failure $\mapsto 0$
- Training set: Movies $0 - 200$, out of 290
- Model: success $= \beta_1\theta_1 + \beta_2\theta_2 + \cdots + \beta_{20}\theta_{20}$

Approach 2: Logistic Regression

- IMDB score $>$ 7.4 $\mapsto$ successful movie $\mapsto$ 1
- IMDB score $<$ 7.4 $\mapsto$ failure $\mapsto$ 0
- Training set: Movies $0-200$, out of 290
- Model: success $= \beta_1\theta_1 + \beta_2\theta_2 + \cdots + \beta_{20}\theta_{20}$

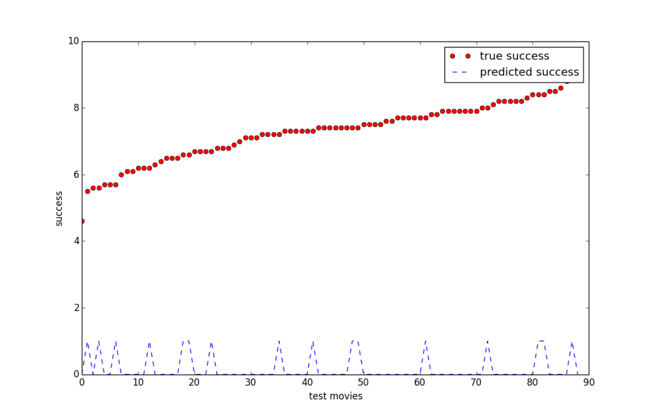
Approach 2: Logistic Regression

- IMDB score $> 7.4 \mapsto$ successful movie $\mapsto 1$
- IMDB score $< 7.4 \mapsto$ failure $\mapsto 0$
- Training set: Movies $0 - 200$, out of 290
- Model: $\text{success} = \beta_1\theta_1 + \beta_2\theta_2 + \cdots + \beta_{20}\theta_{20}$

Approach 2: Logistic Regression

- IMDB score $> 7.4 \mapsto$ successful movie $\mapsto 1$
- IMDB score $< 7.4 \mapsto$ failure $\mapsto 0$
- Training set: Movies $0 - 200$, out of 290
- Model: success $= \beta_1\theta_1 + \beta_2\theta_2 + \cdots + \beta_{20}\theta_{20}$

Predicted vs. actual rating, Logistic Regression



Further approaches

We also tried

- 10-fold cross validation for the logistic regression
- Support Vector Machine
- Support Vector Regression
- Non-Linear Regression (degree 2)

Without achieving better results.

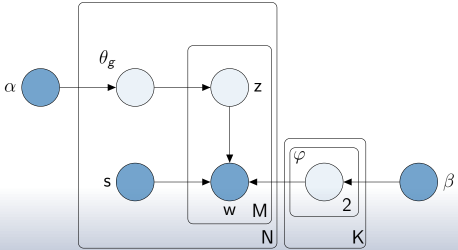
Modified Approach

- We need to modify our model:
- Include binary success variable in the generative model.

Modified Approach

- We need to modify our model:
- Include binary success variable in the generative model.

Modified generative model (fixed genre)



Still To Do

- Work with altered model
- Refine other approaches:
- Binarize θ_i in linear regression (using a threshold, or top n languages)
- Use $\log \theta_i$ in both linear and logistic regression
- Include a buffer zone between success/failure movies