Using scores to improve language modelling of movie plot summaries

J. Saez Gómez, R. v Heijden, F. Stablum

Problem formulation

TOTTILIALI

Approac

Datasat

Results

Discussior

hallenges

Using scores to improve language modelling of movie plot summaries

Jorge Sáez Gómez Roelof van der Heijden Francesco Stablum

Universiteit van Amsterdam

December 16, 2014

Presentation outline

Using scores to improve language modelling of movie plot summaries

J. Sáez Gómez, R. v Heijden, F. Stablum

Problem formulation

Approac

Дрргоас

Results

Discussion

hallenge

- 1 Problem formulation
- 2 Models
- 3 Approach
- 4 Dataset
- 5 Results
- 6 Discussion
- 7 Challenges

Problem formulation

Using scores to improve language modelling of movie plot summaries

Gómez, R. Heijden, F Stablum

Problem formulation

Models

Approa

Datasa

Results

Discussion

hallenges

- Is there any correlation between the score of a movie and the contents of its script?
- Can we use the score to better model a movie corpus?

Models

Using scores to improve language modelling of movie plot summaries

Gómez, R. vo Heijden, F. Stablum

Problem formulation

Models

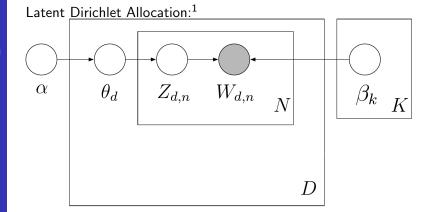
Approac

_

Results

Discussion

hallan a



 $^{^{1}\}text{Image taken from the paper "Supervised topic models" by David M. Blei and Jon D. McAuliffe (2007) <math display="inline">\c \sim$

Models

Using scores to improve language modelling of movie plot summaries

J. Sáez Gómez, R. vo Heijden, F. Stablum

Problem formulation

Models

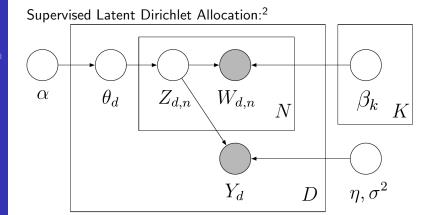
Approa

Datas

Results

Discussion

hallan a



 $^{^2}$ Image taken from the paper "Supervised topic models" by David M. Blei and Jon D. McAuliffe (2007) $\ \ \bigcirc$

Approach

Using scores to improve language modelling of movie plot summaries

J. Sáez Gómez, R. v Heijden, F. Stablum

Problem formulati

Models

Approach

Datase

Results

Discussio

halleng

Our collapsed Gibbs sampler:

$$p(z_{di} = k \mid Z^{\setminus i}, S, W, \alpha, \beta, \eta, \sigma) \propto \left[\prod_{k'} \frac{\prod_{w} \Gamma(N_{k'w}^{\setminus i} + \mathbb{I}(k' = k \land w = w_{di}) + \beta)}{\Gamma(N_{k'}^{\setminus i} + \mathbb{I}(k' = k) + W\beta)} \right] \times \mathcal{N} \left(s_{d} \mid \eta^{T} \cdot \frac{N_{dk'}^{\setminus i} + \mathbb{I}(k' = k)}{N_{d}}, \sigma \right) \prod_{k'} \Gamma(N_{dk'}^{\setminus i} + \mathbb{I}(k' = k) + \alpha)$$
Movie score term

Better implemented in log-space probabilities to avoid

numerical problems.

Approach

Using scores to improve language modelling of movie plot summaries

J. Sáez Gómez, R. v Heijden, F. Stablum

formulation

Models

Approach

Datas

Discussion

hallenges

Estimating the global score hyperparameter η :

$$\eta_k^{new} \leftarrow (1 - \gamma)\eta_k^{old} + \gamma \frac{\sum_d \frac{N_{dk}}{N_d} \left(s_d - \sum_{k' \neq k} \eta_{k'}^{old} \frac{N_{dk'}}{N_d}\right)}{\sum_d \left(\frac{N_{dk}}{N_d}\right)^2 + \varepsilon}$$

Where:

- $1 \gg \gamma > 0$ in order for the previous series to converge.
- $1 \gg \varepsilon > 0$ is a smoothing constant.

Dataset

Using scores to improve language modelling of movie plot summaries

J. Sáez Gómez, R. vo Heijden, F. Stablum

formulati

Models

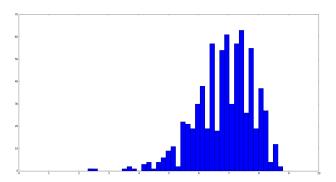
Dataset

Poculto

Discussior

hallenge

- We made scripts to crawl http://www.imsdb.com/ for movie scripts and then search http://www.imdb.com/ for movie scores and plot summaries.
- We got a database with ≈ 700 movies.
- Movie score distribution (from 0 to 10):



Dataset

Using scores to improve language modelling of movie plot summaries

J. Saez Gómez, R. v Heijden, F. Stablum

formulat

Models

Approac

Dataset

.. ..

- lacktriangleright Tokenization o stemming o pruning
- We prune words appearing only on a single movie (avoids overfitting) or within a stop list.
- Total number of tokens $\approx 12.7 \cdot 10^6$
- Number of unique tokens \approx 35000
- Average number of tokens within a movie summary ≈ 75
- \blacksquare Average number of tokens within a movie script ≈ 18000

Results

Using scores to improve language modelling of movie plot summaries

J. Sáez Gómez, R. v Heijden, F. Stablum

formulat

Models

Approx

Datas

Results

Discussion

Challenge

- Initial selection of 120 movies (100 training / 20 testing) with balanced scores.
- Perplexity measure:

LDA topics		25	50	100
Using scores?		4128		
	Yes	4005	5503	7082

Inverse accuracy measure:

LDA topics		25	50	100
Using scores?	No	845	1318	2049
	Yes	805	1372	2067

Results

Using scores to improve language modelling of movie plot summaries

J. Sáez Gómez, R. v Heijden, F. Stablum

Problem formulati

Models

прргос

Results

Discussion

Challenge

- Using 300 movies (250 training / 50 testing).
- Results averaged across 3 MCMC chains.
- Perplexity measure:

LDA topics		25
Hainer agares?	No	3237
Using scores?	Yes	2955

Inverse accuracy measure:

LDA topics		25
Using scores?	No	597
	Yes	571

Discussion

Using scores to improve language modelling of movie plot summaries

J. Sáez Gómez, R. v Heijden, F. Stablum

Problem formulation

. . . .

Appro

Datas

Result

Discussion

hallenges

- Use of few movies \Rightarrow the model starts overfitting with 50-100 topics already.
- Slight predictive improvement if using movie scores, but it is not significant.

Challenges

Using scores to improve language modelling of movie plot summaries

J. Sáez Gómez, R. v Heijden, F. Stablum

formulation

NA - Julia

Approac

....

Results

Discussion

Challenges

- Improve speed of the collapsed Gibbs sampler.
- Use the movie scripts instead of the movie summaries.
- Use both the movie scripts and summaries.
- Incorporate more information into the model, such as the movie genre.