Using scores to improve language modelling of movie plot summaries

Jorge Sáez Gómez
Roelof van der Heijden
Francesco Stablum

Universiteit van Amsterdam

December 10, 2014

Presentation outline

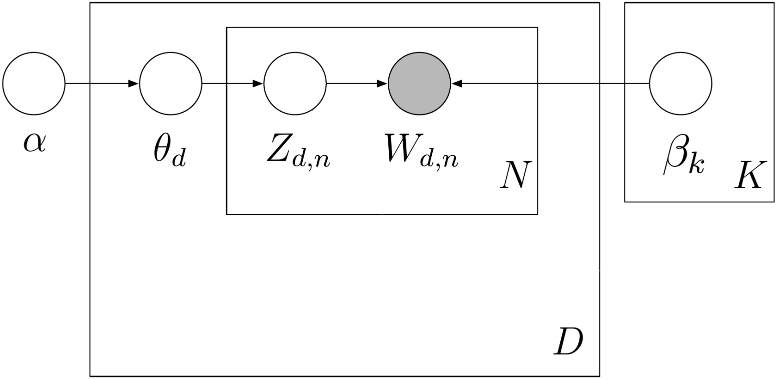
1. Problem formulation
2. Models
3. Approach
4. Dataset
5. Results
6. Discussion
7. Challenges

Problem formulation

- 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

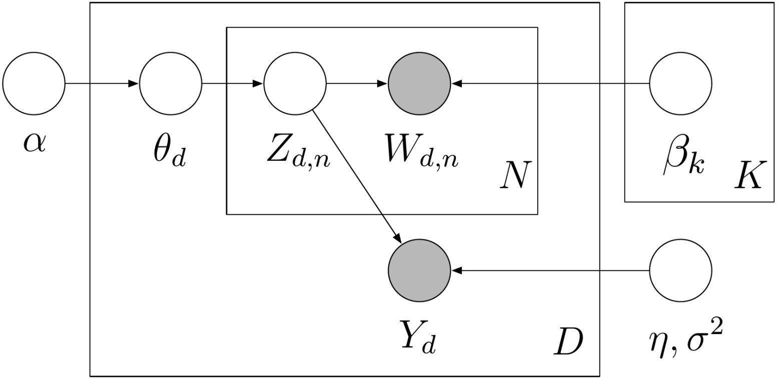
Latent Dirichlet Allocation:[1]



[1] Image taken from the paper "Supervised topic models" by David M. Blei and Jon D. McAuliffe (2007)

Models

Supervised Latent Dirichlet Allocation:[2]



[2] Image taken from the paper "Supervised topic models" by David M. Blei and Jon D. McAuliffe (2007)

Approach

Our collapsed Gibbs sampler:

$$p(z_{di} = k \mid Z^{\setminus i}, S, W, \alpha, \beta, \eta, \sigma) \propto \left[\prod_{k'} \frac{\prod_w \Gamma(N_{k'w}^{\setminus i} + \mathbb{I}(k' = k \wedge w = w_{di}) + \beta)}{\Gamma(N_{k'}^{\setminus i} + \mathbb{I}(k' = k) + W\beta)} \right] \times \underbrace{\mathcal{N}\left(s_d \,\middle|\, \eta^T \cdot \frac{N_{dk'}^{\setminus i} + \mathbb{I}(k' = k)}{N_d}, \sigma \right)}_{\text{Movie score term}} \prod_{k'} \Gamma(N_{dk'}^{\setminus i} + \mathbb{I}(k' = k) + \alpha)$$

Better implemented in log-space probabilities to avoid numerical problems.

Approach

Estimating the global score hyperparameter η:

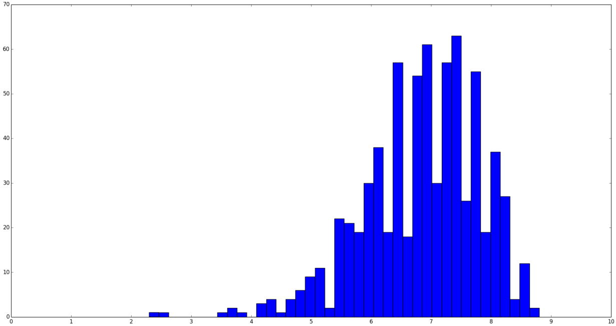
$$\eta_k^{new} \leftarrow (1-\gamma)\eta_k^{old} + \gamma \frac{\sum_d \frac{N_{dk}}{N_d}\left(s_d - \sum_{k'\neq k} \eta_{k'}^{old} \frac{N_{dk'}}{N_d}\right)}{\sum_d \left(\frac{N_{dk}}{N_d}\right)^2 + \varepsilon}$$

Where:

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

Dataset

- 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):

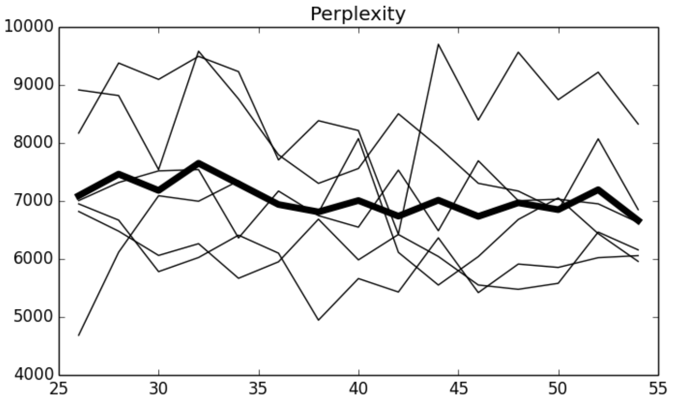


Dataset

- Tokenization → stemming → 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 ≈ 35000
- Average number of tokens within a movie summary ≈ 75
- Average number of tokens within a movie script ≈ 18000

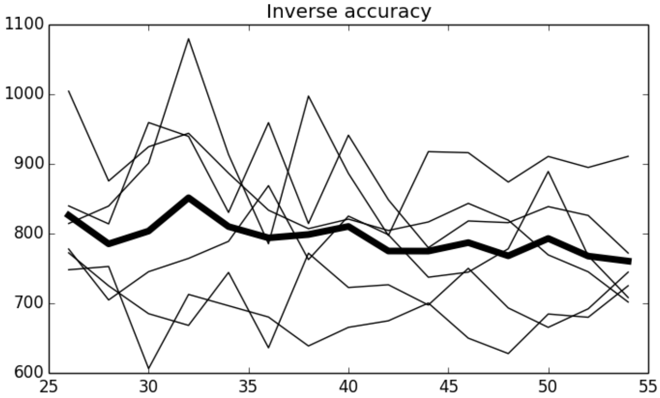
Results

- Initial selection of 30 movies (20 training / 10 testing) with balanced scores.
- Using 10 topics and no scores.



Results

- Initial selection of 30 movies (20 training / 10 testing) with balanced scores.
- Using 10 topics and no scores.



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.