1 Introduction

In recent years, several successes have been booked for applying semantic analysis on user comments of movies. In this report we use those same techniques, but apply them to plot summaries of movies to try and estimate the score these movies are rated with on the popular online movie database IMDb. These user-written texts roughly describe what happens in a particular movie and may therefore be an indication of the score, which is also calculated from user submitted data.

This project is part of the Natural Language Processing course of the UvA from Fall 2014.

2 Models

In this section we describe the extended topic based model we used. It is taken from [**?**].

The generative version of LDA is as follows:

1. Draw topic proportions $\theta \mid \alpha \sim \mathrm{Dir}(\alpha)$.
2. For each word:
   (a) Draw topic assignment $z_n \mid \theta \sim \mathrm{Mult}(\theta)$.
   (b) Draw word $w_n \mid z_n, \beta_{1:K} \sim \mathrm{Mult}(\beta_{zn})$.

Its graphical representation can be seen in Figure 2.

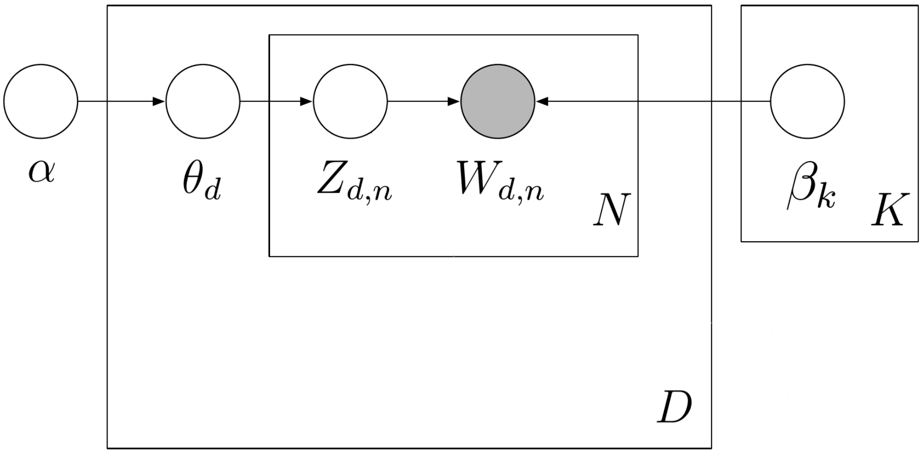

Figure 1: A graphical representation of traditional LDA model.

We however, use the an extended version of LDA, which makes use of the given scores. Because of this, a third step is added to the generative process:

1. Draw topic proportions $\theta \mid \alpha \sim \mathrm{Dir}(\alpha)$.
2. For each word:
    (a) Draw topic assignment $z_n \mid \theta \sim \mathrm{Mult}(\theta)$.
    (b) Draw word $w_n \mid z_n, \beta_{1:K} \sim \mathrm{Mult}(\beta_{zn})$.
3. Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim \mathcal{N}(\eta^\top \bar{z}, \sigma^2)$.

This version can be called supervised LDA or SLDA. Its graphical representation can be seen in Figure 2.

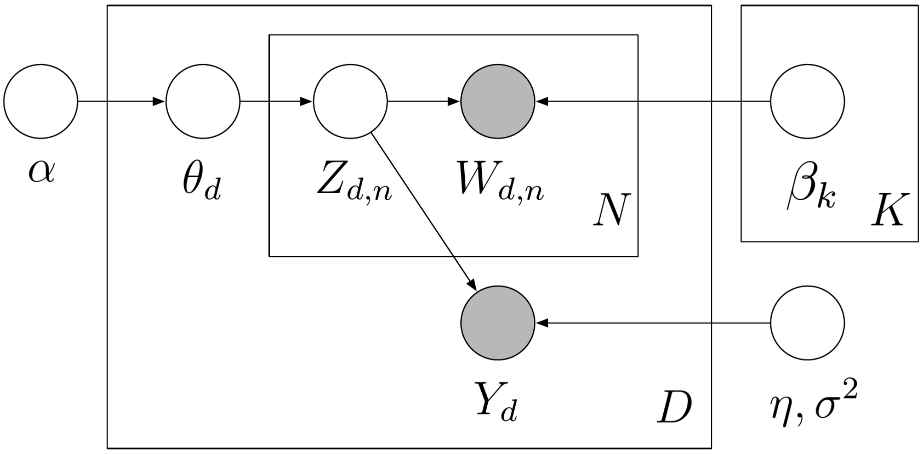

Figure 2: A graphical representation of our modified LDA model.

3 Approach

(I took this straight from the photo Francesco sent.)

$$P(\theta, s, z, p, W \mid \alpha, \beta) = \prod_d \mathrm{Dir}(\varphi_d \mid \beta) \tag{1}$$

$$\left[\prod_d \mathrm{Dir}(\theta_d \mid \alpha) \prod_d \mathrm{Mult}(\varphi_d \mid \theta_d)\, \mathrm{Mult}(w_d \mid \varphi_d)\right]$$

$$\int_{\varphi_0} \int_{\varphi_1} \dots \int_{\varphi_k} P(\theta, s, z, \varphi, W \mid \alpha, \beta, \eta, \sigma)$$

$$= \left[\prod_k \mathrm{Dir}(\varphi_k \mid \beta)\right] \left[\prod_d \mathrm{Dir}(\theta_d \mid \alpha) \mathcal{N}(\eta^\top \bar{z}_d, \sigma) \prod_i^{N_d} \mathrm{Mult}(z)\right]$$

$$\times \left[\prod_d \prod_i^{N_d} \mathrm{Mult}(w_{di} \mid \varphi_{z_d})\right] \to \prod_d \prod_w \prod_k \left[\mathrm{Mult}(w \mid \varphi_k)^{N_{dk}}\right]$$

$$P(\theta, s, z, w \mid \alpha, \beta, \eta, \sigma) = \left[\prod_k \int_{\varphi_k} \mathrm{Dir}(\varphi_k \mid \beta) \prod_d \prod_w \left[\mathrm{Mult}(w \mid \varphi_k)^{N_{dk}}\right]\right]$$

$$= \left[\prod_d \mathrm{Dir}(\theta_d \mid \alpha) \mathcal{N}(s_d \mid \eta^\top \bar{z}_d, \sigma) \prod_i^N \mathrm{Mult}(z_{di} \mid \theta_d)\right]$$

$$\times \left[\prod_k \frac{\Gamma(\beta)\Gamma(W\beta)}{\Gamma(N_k + W\beta)} \prod_w \frac{\Gamma(N_{kw} + \beta)}{\Gamma(\beta)}\right]$$

next page

$$P(s, z, w \mid \alpha, \beta, \eta, \sigma) =$$

$$\left[\prod_k \frac{W\beta}{\Gamma(\beta)^W \Gamma(N_k + W\beta)} \prod_w \Gamma(N_{kw} + \beta)\right]$$

$$\times \left[\prod_d \mathcal{N}(s_d \mid \eta^\top \bar{z}_d, \sigma) \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K \Gamma(N_d + K\alpha)}\right.$$

$$\left.\times \prod_k \Gamma(N_{dk} + \alpha)\right]$$

$$P(z \mid s, W, \dots) \overset{\text{Bayes}}{=} \frac{P(s, w \mid z, \dots) P(z \mid \dots)}{P(s, w \mid \dots)}$$

$$\bar{z}_d = \frac{N_{dk}}{N_d}$$

$$P(z_i = k \mid z_{-i}, s, w, \dots) \propto P(z_i = k \mid z_{-i}, s, w)$$

$$= (\textit{kind of}) \left[\prod_{k'} \frac{1}{\Gamma(N_k + W\beta)} \prod_w \Gamma(N_{kw} + \beta)\right]$$

$$\times \mathcal{N}\left(s_d \mid \eta^\top \frac{N_{dk}}{N_d}, \sigma\right) \frac{1}{\Gamma(N_d + K\alpha)} \prod_{k'} \Gamma(N_{dk} + \alpha)$$

and we conclude Collapsed Segmented LDA (CSLDA):

$$P(S, Z, W \mid \alpha, \beta, \eta, \sigma) = \left[\prod_k \frac{\Gamma(W\beta)}{\Gamma(\beta)^W \cdot \Gamma(N_k + W\beta)} \prod_w \Gamma(N_{kw} + \beta)\right]$$

$$\times \left[\prod_d \mathcal{N}\left(s_d \,\middle|\, \eta^T \cdot \frac{N_{dk}}{N_d}, \sigma\right) \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K \cdot \Gamma(N_d + K\alpha)} \prod_k \Gamma(N_{dk} + \alpha)\right]$$

Where we used $\frac{N_{dk}}{N_d} \equiv \bar{Z}_d$

3.1 Transformation to log-space

Now we transform into log-space:

$$\log P(S, Z, W \mid \alpha, \beta, \eta, \sigma) =$$

$$\left[\sum_k \log\Gamma(W\beta) - W\log\Gamma(\beta) - \log\Gamma(N_k + W\beta) + \sum_w \log\Gamma(N_{kw} + \beta)\right] +$$

$$\left[\sum_d \underbrace{-\log\sigma - \frac{1}{2}\log(2\pi) - \frac{\left(s_d - \eta^T \cdot \frac{N_{dk}}{N_d}\right)^2}{2\sigma^2}}_{\text{Normal distribution}} + \right.$$

$$\left. \log\Gamma(K\alpha) - K\log\Gamma(\alpha) - \log\Gamma(N_d + K\alpha) + \sum_k \log\Gamma(N_{dk} + \alpha)\right]$$

3.2 Estimating Gamma-function

The logarithm of the gamma function can be rewritten as follows [**?**]:

$$\log\Gamma(z) = -\gamma z - \log z + \sum_{j=1}^{\infty}\left[\frac{z}{j} - \log(1 + \frac{z}{j})\right] \tag{2}$$

where γ is the Euler-Mascheroni constant.We apply this to $\sum_k \sum_w \log\Gamma(N_{kw} + \beta)$, which then becomes:

$$= \sum_{k,w} -\gamma(N_{kw} + \beta) - \log(N_{kw} + \beta) + \sum_{j=1}^{\infty} \frac{N_{kw} + \beta}{j} - \log\left(1 + \frac{N_{kw} + \beta}{j}\right)$$

$$= -\gamma(N + KW\beta) - \sum_{k,w} \log(N_{kw} + \beta) - \sum_{j=1}^{\infty} \frac{N_{kw} + \beta}{j} - \log\left(\frac{N_{kw} + \beta + j}{j}\right)$$

Note that the term $-\gamma(N + KW\beta)$ serves as a normalisation constant for this dataset. Since we do not need the exact probabilities but only the proportional probabilities during the algorithms execution, we can discard those terms.

$$\Rightarrow -\sum_{k,w} \log(N_{kw} + \beta) - \sum_{j=1}^{\infty} \frac{N_{kw} + \beta}{j} - \log\left(\frac{N_{kw} + \beta + j}{j}\right)$$

$$= -\sum_{k,w} \log(N_{kw} + \beta) - \sum_{j=1}^{\infty} \frac{N_{kw} + \beta}{j} - \log\left(N_{kw} + \beta + j\right) + \log(j))$$

$$= -\sum_{k,w} \log(N_{kw} + \beta) - \sum_{j=1}^{\infty} \left(\frac{N_{kw} + \beta}{j} + \log(j)\right) + \sum_{j=1}^{\infty} \log(N_{kw} + \beta + j)$$

$$= \sum_{j=1}^{\infty} \left(\frac{N + KW\beta}{j} + KW\log(j)\right) - \sum_{k,w} \log(N_{kw} + \beta) + \sum_{j=1}^{\infty} \log(N_{kw} + \beta + j)$$

$$= \sum_{j=1}^{\infty} \left(\frac{N + KW\beta}{j} + KW\log(j)\right) - \sum_{k,w} \sum_{j=0}^{\infty} \log(N_{kw} + \beta + j)$$

Again, $\sum_{j=1}^{\infty} \frac{N+KW\beta}{j} + KW\log(j)$ is a constant for this dataset, so we can discard it. This results in the following proportionality:

$$\sum_{k,w} \log\Gamma(N_{kw} + \beta) \propto -\sum_{k,w}\sum_{j=0}^{\infty} \log(N_{kw} + \beta + j) \tag{3}$$

Using similar steps, we can also simplify

$$\sum_{k} \log\Gamma(N_{k} + W\beta) \propto -\sum_{k}\sum_{j=0}^{\infty} \log(N_{k} + W\beta + j) \tag{4}$$

$$\sum_{k,d} \log\Gamma(N_{dk} + \alpha) \propto -\sum_{k,d}\sum_{j=0}^{\infty} \log(N_{dk} + \alpha + j) \tag{5}$$

$$\sum_{d} \log\Gamma(N_{d} + K\alpha) \propto -\sum_{d}\sum_{j=0}^{\infty} \log(N_{d} + K\alpha + j) \tag{6}$$

Putting these together gives us the following proportionality:

$$\begin{aligned} \log P(S, Z, W \mid \alpha, \beta, \eta, \sigma) \propto \frac{1}{2\sigma^2} \sum_d \left(s_d - \eta^T \cdot \frac{N_{dk}}{N_d} \right)^2 - \\ \sum_{j=0}^{\infty}\sum_{d} \log(N_d + K\alpha + j) - \sum_{j=0}^{\infty}\sum_{k,d} \log(N_{dk} + \alpha + j) - \\ \sum_{j=0}^{\infty}\sum_{k} \log(N_k + W\beta + j) - \sum_{j=0}^{\infty}\sum_{k,w} \log(N_{kw} + \beta + j) \end{aligned}$$

4 Dataset

5 Experiments

6 Results

7 Discussion

Maximum a posteriori (MAP) estimate for the η hyperparameter:

$$\nabla_{\eta_k} \log p(S, Z, W|\alpha, \beta, \eta, \sigma) = \sum_d \frac{\frac{N_{dk}}{N_d}\left(s_d - \eta^T \frac{N_{d\cdot}}{N_d}\right)}{\sigma^2} =$$

$$= \sum_d \frac{s_d \frac{N_{dk}}{N_d}}{\sigma^2} - \sum_d \frac{\frac{N_{dk}}{N_d}\left(\eta^T \frac{N_{d\cdot}}{N_d}\right)}{\sigma^2} = 0$$

$$\Rightarrow \sum_d s_d \frac{N_{dk}}{N_d} = \sum_d \frac{N_{dk}}{N_d} \left(\sum_{k'} \eta_{k'} \frac{N_{dk'}}{N_d} \right) = \sum_d \frac{N_{dk}}{N_d} \left(\eta_k \frac{N_{dk}}{N_d} + \sum_{k' \neq k} \eta_{k'} \frac{N_{dk'}}{N_d} \right)$$

$$\Rightarrow \sum_d s_d \frac{N_{dk}}{N_d} = \eta_k \sum_d \left(\frac{N_{dk}}{N_d} \right)^2 + \sum_d \left(\frac{N_{dk}}{N_d} \sum_{k' \neq k} \eta_{k'} \frac{N_{dk'}}{N_d} \right)$$

$$\Rightarrow \sum_d \left(s_d \frac{N_{dk}}{N_d} - \frac{N_{dk}}{N_d} \sum_{k' \neq k} \eta_{k'} \frac{N_{dk'}}{N_d} \right) = \eta_k \sum_d \left(\frac{N_{dk}}{N_d} \right)^2$$

$$\Rightarrow \sum_d \frac{N_{dk}}{N_d} \left(s_d - \sum_{k' \neq k} \eta_{k'} \frac{N_{dk'}}{N_d} \right) = \eta_k \sum_d \left(\frac{N_{dk}}{N_d} \right)^2$$

$$\Rightarrow \eta_k = \frac{\sum_d \frac{N_{dk}}{N_d} \left(s_d - \sum_{k' \neq k} \eta_{k'} \frac{N_{dk'}}{N_d} \right)}{\sum_d \left(\frac{N_{dk}}{N_d} \right)^2}$$

Trying to apply the previous formula as an update rule for η does not converge. Instead, the following update can be used:

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

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

Gibbs sampler:

$$\begin{aligned} p(z_{di} = k | Z^{\setminus i}, S, W, \alpha, \beta, \eta, \sigma) &\propto p(z_{di} = k, Z_{-i}, S, W, \alpha, \beta, \eta, \sigma) \\ &\propto \left[\prod_{k'} \frac{\prod_w \Gamma(N_{k'w}^{\setminus i} + 1(k' = k \wedge w = w_{di}) + \beta)}{\Gamma(N_{k'}^{\setminus i} + 1(k' = k) + W\beta)} \right] \times \\ \mathcal{N}\left(s_d \,\middle|\, \eta^T \cdot \frac{N_{dk'}^{\setminus i} + 1(k' = k)}{N_d}, \sigma \right) &\prod_{k'} \Gamma(N_{dk'}^{\setminus i} + 1(k' = k) + \alpha) \end{aligned}$$