Automated Survey Text Analysis - Supervised Latent Dirichlet Allocation (sLDA)

Christine P. Chai (chrchai@microsoft.com)

Core Services Engineering Operations (CSEO), Microsoft

Disclaimer

The research was conducted while the author was a PhD student in statistical science at Duke University. The views expressed in this e-poster are those of the author and do not reflect the view of Microsoft.

Overview

Open-ended questions are becoming more common in surveys, due to the diverse responses they can capture. However, the analysis of survey text is often conducted manually, which can be expensive and prone to subjectivity. Therefore, we would like to automatically analyze text and numerical data using the supervised latent Dirichlet allocation (sLDA), a topic modeling approach that assigns each word a probability distribution of topics. The example we used is an employee satisfaction survey, and each record contains a numerical rating along with a free text response as the reason. Then the sLDA algorithm selects key words of each rating as a topic, and outputs the corresponding credible intervals. Since the R package `lda` is available for this approach, using sLDA to identify topics for each rating is a start for automated survey text analysis, with little technical knowledge required for implementation.

Employee Satisfaction Survey

The data contain 530 employee ratings and text comments about their work. The ratings are from 1-10, with 1 the least satisfied and 10 the most. Most ratings are between 5 and 9, and each comment has 23.54 words on average.

sLDA Algorithm

sLDA is a Bayesian data generative process [2]. For each document D_d

- Draw topic proportions $\theta_d|\alpha \sim \text{Dirichlet}(\alpha)$
- For each word $W_{d,n}$
  - Draw topic assignment $Z_{d,n}|\theta_d \sim \text{Multinomial}(\theta_d)$
  - Draw word $W_{d,n}|Z_{d,n}, \beta_{1:K} \sim \text{Multinomial}(\beta_{Z_{d,n}})$
- Draw response variable Y_d
  - $Y_d|Z_{d,1:N_d}, \eta, \sigma^2 \sim N(\eta\bar{Z}_d, \sigma^2)$
  - $\bar{Z}_d = (1/N_d)\sum_{n=1}^{N_d} Z_{d,n}$

Transform $Y_d \in \mathbb{R}$ to K topics (scores 1-10)

- $-\infty = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_{K-1} < \tau_K = \infty$
- In the kth category, $\tau_{k-1} \leq Y_d < \tau_k$

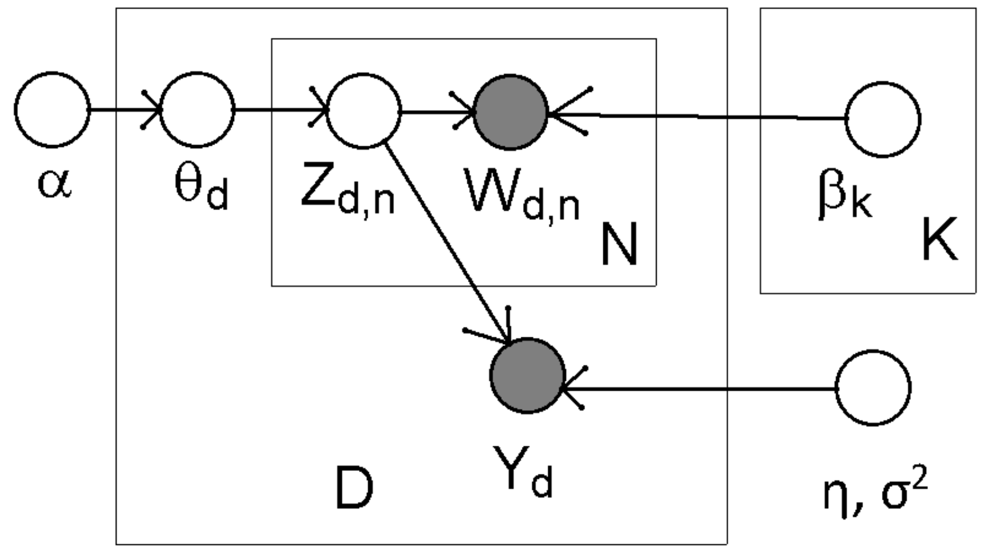

Figure 1: Plate diagram for sLDA [4]

Implementation and Results

- R package `lda`: Sample code in `demo(slda)`
- R function `top.topic.words` [1]
  - Selects five words for each topic (rating)
  - Based on the posterior $P($ word $j|$ topic $i,$ data $)$
- R function `slda.predict` for new comments

Table 1: Selected words (tokens) for each topic (rating)

Rating (Topic)	Selected Words
10	challeng opportun learn dai new
9	team work great staff make
8	role feel current perform career
7	job work **enjoi** too project
6	manag work happi last month
5	work get **enjoi** chang need
4	compani project skill resourc engin
3	life balanc compani work peopl
2	lack interest differ **enjoi** requir
1	time hour lot week take

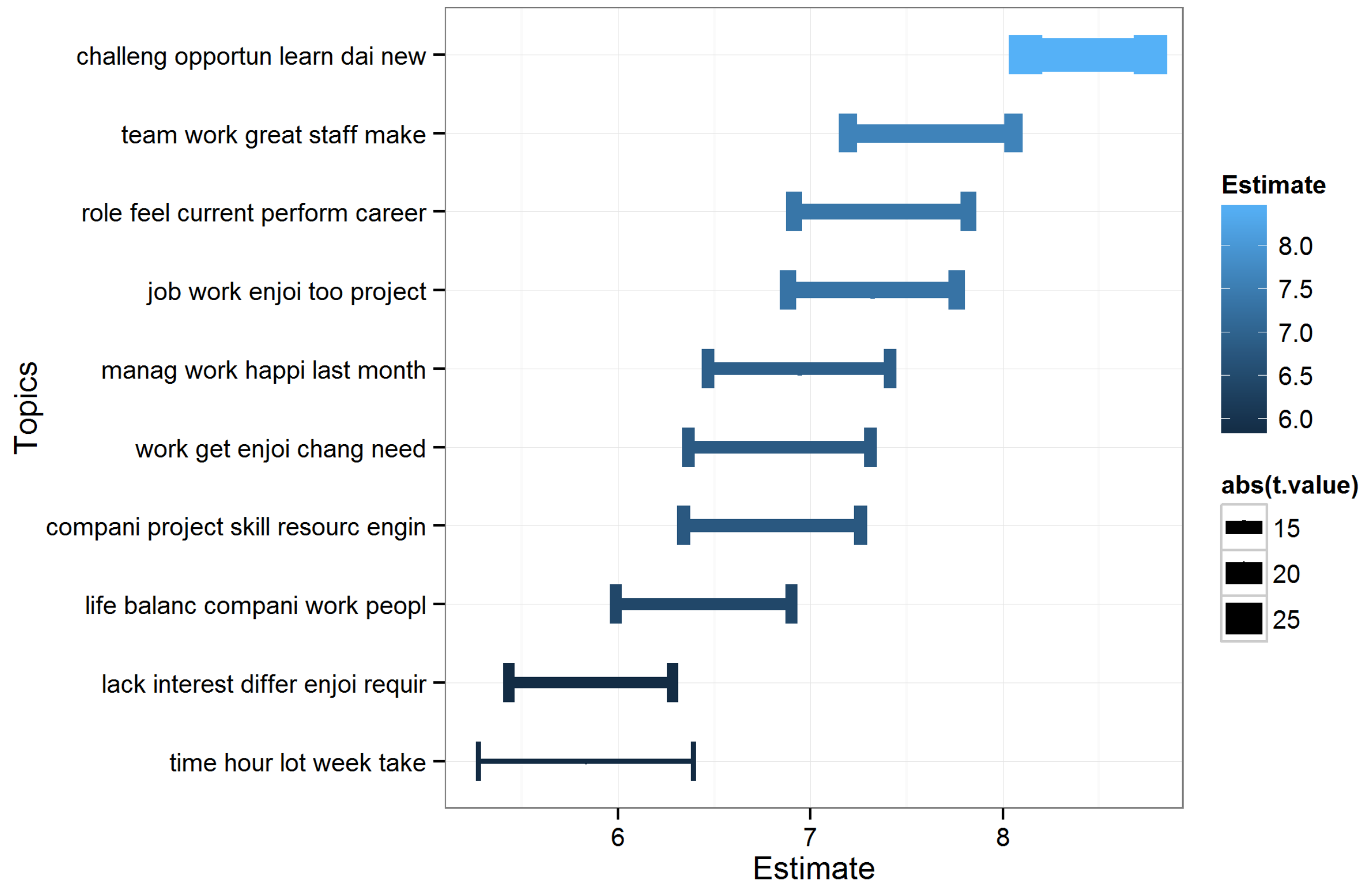

Figure 2: 68% credible intervals for each topic [3]

Discussion and Conclusion

sLDA is a start for automated survey text analysis, which would be helpful in analyzing large amounts of text responses. A potential application is to detect errors by comparing the actual rating with the comment's estimated rating. Currently, we would like to further improve the results before recommending full migration to a supervised automation approach. For example, advanced statistical methods are needed to narrow down and calibrate the credible intervals.

Acknowledgments

The data are from Nick Fisher at ValueMetrics, Australia. The author would like to thank David Banks, her PhD advisor at Duke University, for his support on this research project. The author is also grateful for the conference registration and support provided by Microsoft.

References

[1] J. Chang. `lda`: Collapsed Gibbs Sampling Methods for Topic Models, 2015. R package version 1.4.2.

[2] J.D. McAuliffe & D.M. Blei. Supervised Topic Models. Advances in Neural Information Processing Systems, 121-128, 2008.

[3] C.P. Chai. Statistical Issues in Quantifying Text Mining Performance. PhD Dissertation, Duke University, 2017.

[4] C.P. Chai. Text Mining in Survey Data. Survey Practice, 2019.