Sina Weibo as a Corpus for Studying Public Opinions

Christine Kuang, Siqi Wu, and Angie Zhu

Department of Statistics, UC Berkeley

May 3, 2012

Outline

Introduction

- Opinions on microblogging and social networking websites
- Sina Weibo 新浪微博 is the largest microblogging website:
  accounted for 65% of China's microblog market as of December 2011
- Study public opinions using Sina Weibo as a corpus for a given topic

Topic

- Internet censorship in China
- Time sensitive
- Processing is topic-dependent
- Hot topic is preferred
- Chosen topic: Han Han 韩寒

Background

- HAN Han 韩寒 (born 23 September 1982) is a Chinese best-selling author, professional rally driver, and wildly popular blogger
- Published his first novel *Triple Gate* 三重门 at age of 17
- High school dropout

Photograph by Tony Law / Redux. Source:
`http://www.time.com/time/magazine/article/0,9171,1931619,00.html`

Background

- Ghostwriting allegation against Han from January 2012
- FANG Zhouzi 方舟子, a scientific author and anti-fraud crusader, created widespread debate on the internet
- 光明与磊落
- Han received a death threat on April 15, 2012

Data Collection

- Topic searching via API:
  only the latest results are returned
  up to 30 each time
- Collected on April 16 and 17, 2012

Characteristics of Chinese Language

- No explicit delimiter
- Ambiguities in phrases
  - Context ambiguition: e.g., 他好吃
  - Word definition ambiguition: e.g., 打
- Out-of-vocabulary words
- No 1-to-1 correspondence between traditional and simplified Chinese

Characteristics of Sina Weibo Posts

- .

Pre-tagging Processing

- .

Tagging

- Process: tagged 3000 total posts with four categories
- Examples:

  Positive 支持韩寒！ Support Han Han!

  Negative 看到韩寒就恶心。Feel nauseous when I see Han Han.

- Limitations:
  - Subjective responses:

    e.g., "that wasn't too bad"
  - Uncertain tags
    - Quotes
    - Posts without subjects
    - Posts that just mention opposing author

Pre-segmentation Processing

- .

Segmentation

- .

Conjunction Rules

- .

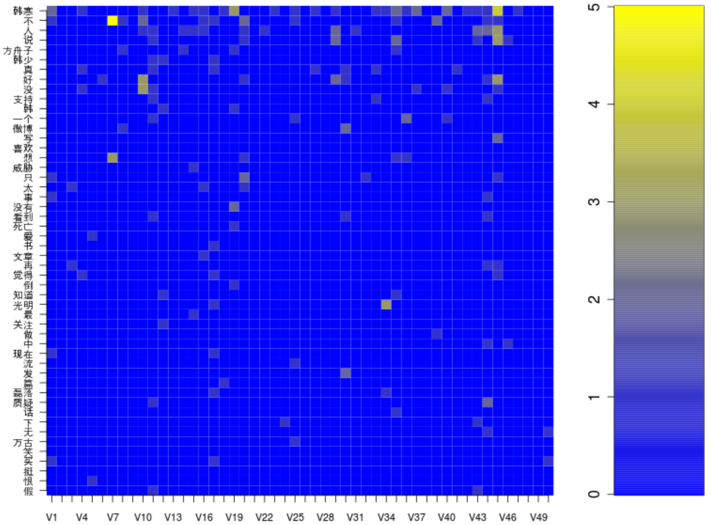
Stop Words and Punctuation Elimination

- .

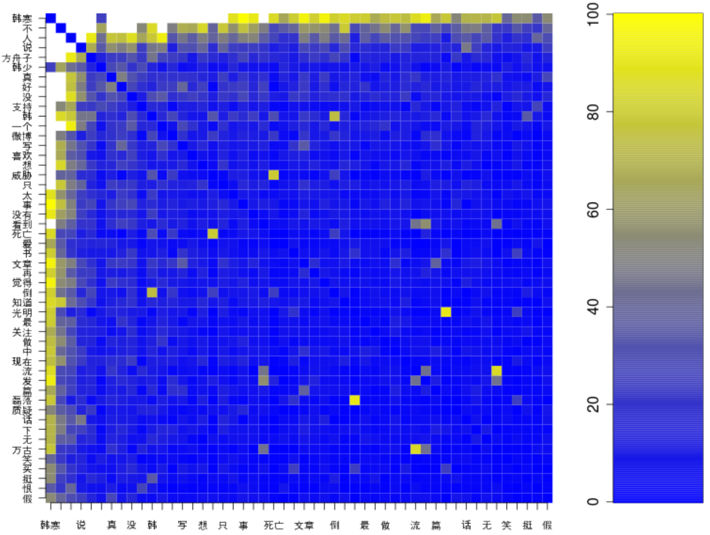
EDA

- ≥ 10

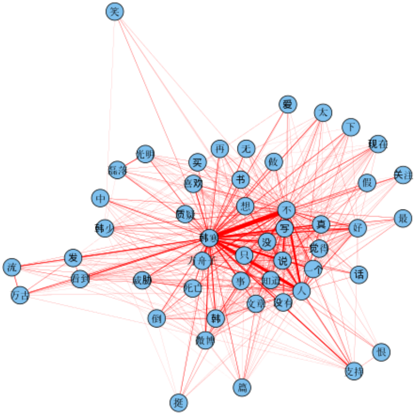
Frequency


韩寒
不
人
说
方舟子
韩少
真
好
没
支持
韩
一个
微博
写
喜欢
想
威胁
只
太
事
没有
看到
死亡
爱
书
文章
再
觉得
倒
知道
光明
最
关注
做
中
现在
流
发
篇
磊落
质疑
话
下
无
万古
笑
买
挺
恨
假
V1 V4 V7 V10 V13 V16 V19 V22 V25 V28 V31 V34 V37 V40 V43 V46 V49
0
1
2
3
4
5

Co-occurrance



Co-occurrance



Graphical LASSO

- Suppose the random vector $x \in R^p$ has a multivariate normal distribution with mean μ and covariance matrix Σ. The density function of x is

$$f(x|\mu, \Sigma) = \frac{1}{\sqrt{2\pi \det(\Sigma)}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}.$$

- Suppose we have IID $N(\nu, \Sigma)$ data $x_1, x_2, ...x_n$ and we want to estimate the inverse covariance matrix Σ^{-1}. The joint likelihood of the data is

$$\begin{aligned} & f(x_1, \ldots, x_n|\mu, \Sigma) \\ & = \frac{1}{(2\pi \det(\Sigma))^{n/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)\right\}. \end{aligned}$$

Graphical LASSO

- Taking logarithm to get the log-likelihood (and ignore constant terms):

$$l(\mu, \Sigma^{-1}) = \frac{1}{2 \log \det (\Sigma)} - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$$

- Do a maximum likelihood estimation (optimize over μ and $S = \Sigma^{-1}$; easy to see that the MLE for μ is $\bar{X}$):

$$\max_{S} \left\{ \frac{1}{2} \log \det (S) - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T S (x_i - \mu) \right\}$$

Graphical LASSO

- Here comes the trace trick $\sum_{i=1}^{n}(x_i-\mu)^T S(x_i-\mu) = \mathbf{Tr}(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T S) = n\mathbf{Tr}(\hat{\Sigma}S)$. We end up with the following optimization problem

$$\max_{S}\left\{\log\det(S) - \mathbf{Tr}\left(\hat{\Sigma}S\right)\right\}$$

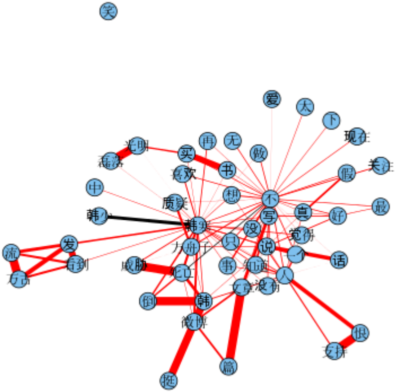
- Banerjee et al. (2007) propose the following optimization problem to recover the sparse structure in a gaussian graphical model

$$\max_{S}\left\{\log\det S - \mathbf{Tr}\left(\hat{\Sigma}S\right) - \lambda||S||_1\right\}$$

where $||S||_1 = \sum_{i=1}\sum_{j=1}|s_{ij}|$.

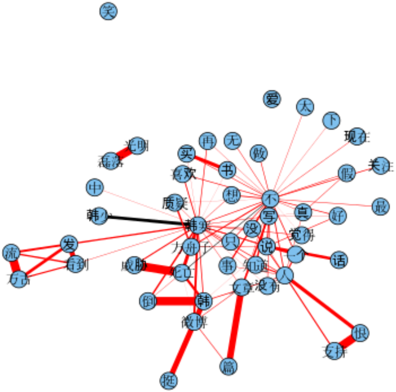
- .

Graphical LASSO



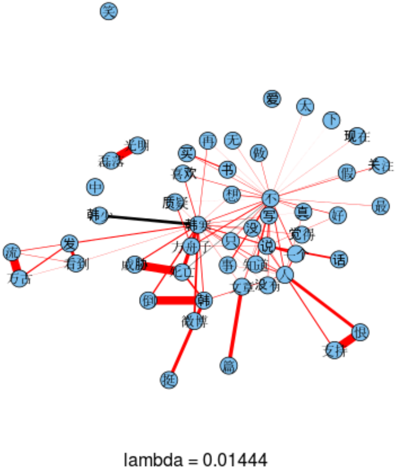
lambda = 0.01

Graphical LASSO

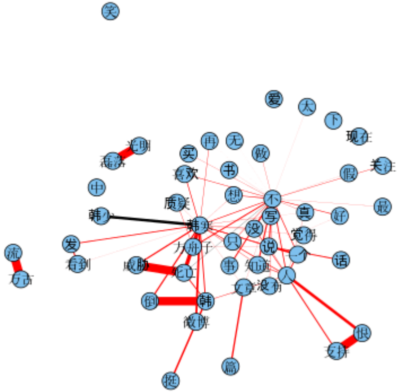


lambda = 0.01222

Graphical LASSO

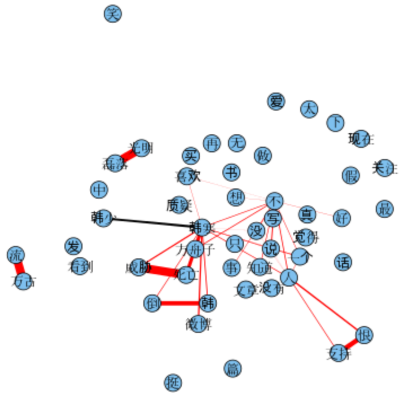


Graphical LASSO



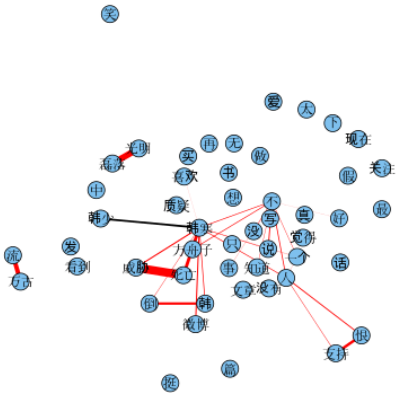
lambda = 0.01667

Graphical LASSO



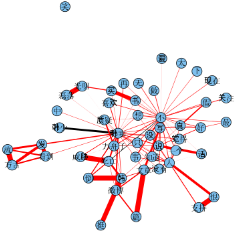
lambda = 0.02556

Graphical LASSO

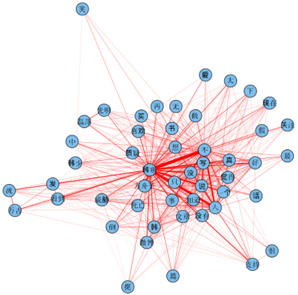


lambda = 0.03

Graphical LASSO



lambda = 0.01



LASSO

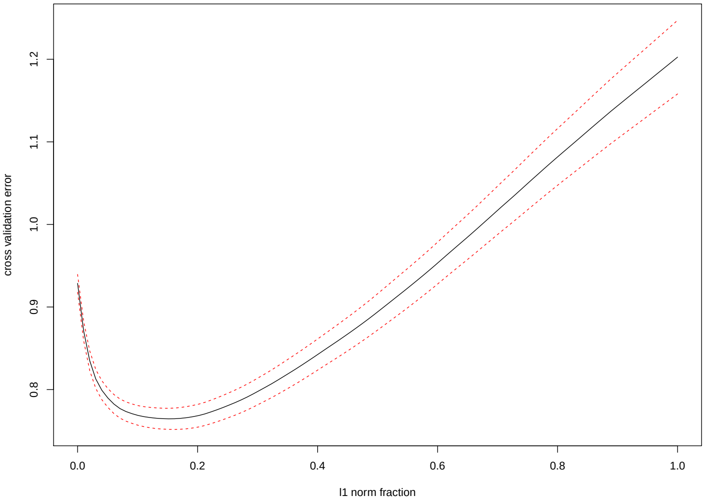
$$\hat{\beta}(\lambda) = \arg\min_{\beta} \frac{1}{2}||y - (\beta_0 + X\beta)||_2^2 + \lambda||\beta||_1$$

- Four models for each category for classification
- General overview of method
- General overview of application to data
    - for 4 categories
    - 10 fold CV
    - Frequency matrix is 3000×795
    - classification error

LASSO Results

- Three different ways to look at coefficients
- Why:
  - Absolute value: most relevant
  - Positive: more likely to be in category
  - Negative: less likely to be in category

LASSO Results: Positive Category



LASSO Results: Positive Category

Word	Absolute Coef.	Word	Positive Coef.	Word	Negative Coef.
加油 (keep going)	0.820	加油 (keep going)	0.820	样子 (manner)	-0.396
韩少 (Master Han)	0.644	韩少 (Master Han)	0.644	恋 (love)	-0.344
成熟 (mature)	0.546	成熟 (mature)	0.546	发表 (announce)	-0.336
顶 (support)	0.533	顶 (support)	0.533	道理 (rational)	-0.336
宽容 (tolerant)	0.518	宽容 (tolerant)	0.518	利益 (benefit)	-0.335

LASSO Results: Negative Category

Word	Absolute Coef.	Word	Positive Coef.	Word	Negative Coef.
讨厌 (hate)	0.481	讨厌 (hate)	0.481	支持 (support)	-0.008
无耻 (shameless)	0.412	无耻 (shameless)	0.412	不 (no)	0.000
恶心 (disgusting)	0.395	恶心 (disgusting)	0.395	人 (people/person)	0.000
骗子 (liar)	0.380	骗子 (liar)	0.380	说 (say)	0.000
扁 (beat up)	0.353	扁 (beat up)	0.353	方舟子 (FangZhouZi)	0.000

l_1-Norm Support Vector Machine

- Support vector machine (Vapnik 1996) is another commonly-used machine learning method to classify data points into two categories. Consider again the linear decision function $f(x) = \beta_0 + \beta x$ and the classifier $Class(x) = \textbf{sign}(f(x))$. To "learn" the parameters, we want the training misclassification rate to be small and the margin of the decision boundary (which can be shown to be $1/||\beta||_2$) to be wide. Hence we consider the following optimization problem (give a 2d illustration here...):

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} (1 - y_i(\beta_0 + \beta^T x_i))_+ + \frac{\lambda}{2}||\beta||_2,$$

where $z_+ = \max(0, z)$ (the function $h(z) = (1 - z)_+$ is also known as the hinge loss function).

l_1-Norm Support Vector Machine

- Similarly, the sparse version of SVM simply replaces the l_2-norm by l_1-norm (which is just another measure of the wideness of the margin):

$$\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{n} (1 - y_i(\beta_0 + \beta^T x_i))_+ + \lambda ||\beta||_1 \right\}.$$

We repeat the same data analysis as we did for the LASSO method. The results are summaried in table XX. For efficiently fitting the sparse svm, we use the matlab package by Fung and Mangasarian (2004). Again, 10-fold cross validations are performed in order to select the "best" λ.

Further Work

- .

- .

- .

- .

- .

- .

- .

- .