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

1. Introduction
2. Processing
3. EDA
4. Classification
   - LASSO
5. Further Work

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

- .

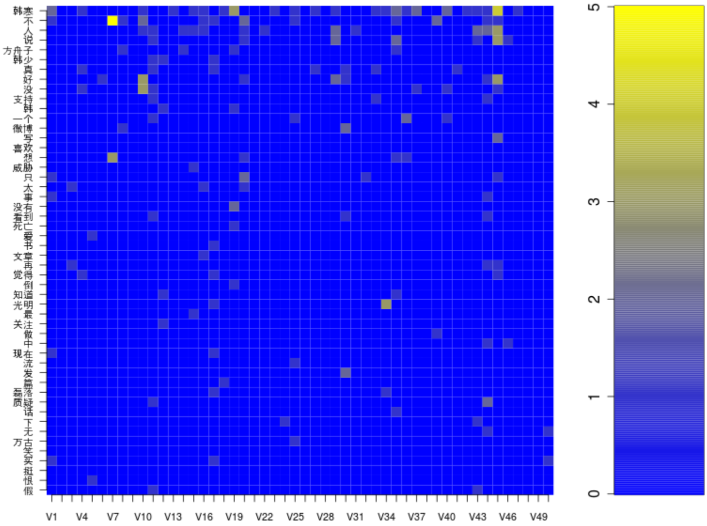
EDA

- ≥ 10

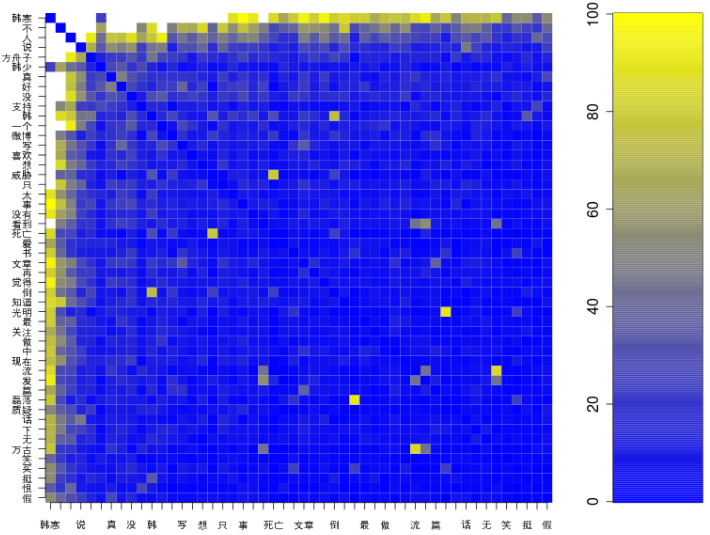
Word frequency

- Extract the word frequency vector x_i from the i-th post
- Construct the word frequency matrix $X = (x_1, ..., x_n)^T$. This will be our design matrix.

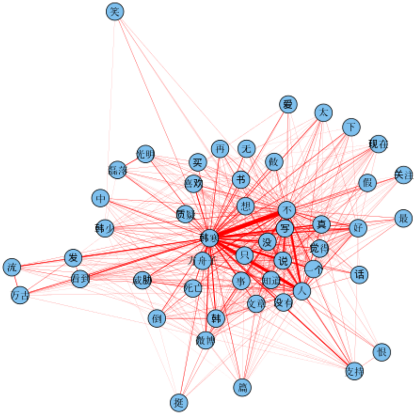
Word frequency visualization: matrix plot



Co-occurrance



Co-occurrance



Sparse graphical models

- Fact: if $x \in R^p$ follows $N(\mu, \Sigma)$, then for $i \neq j$

$$(x_i \perp\!\!\!\perp x_j) \mid \{x_{\text{all but } (i,j)}\} \textbf{ iff } (\Sigma^{-1})_{ij} = 0$$

- This motivates us to estimate Σ^{-1}.
- Let $x_1, x_2, ...x_n$ be IID $N(\nu, \Sigma)$ data. The joint likelihood of the data is

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

Sparse graphical models (cont'd)

- Log-likelihood:

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

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

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

Sparse graphical models (cont'd)

- 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 optimization problem for fitting a Gaussian graphical model:

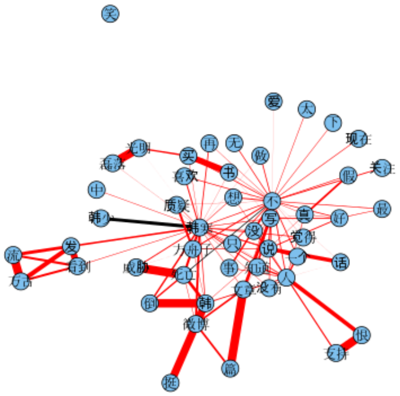
$$\max_S \left\{ \log\det(S) - \mathbf{Tr}\left[\hat{\Sigma}S\right[\right\}$$

- Fitting a sparse Gaussian graphical model:

$$\max_S \left\{ \log\det S - \mathbf{Tr}\left[\hat{\Sigma}S\right[- \lambda||S||_1\right\}$$

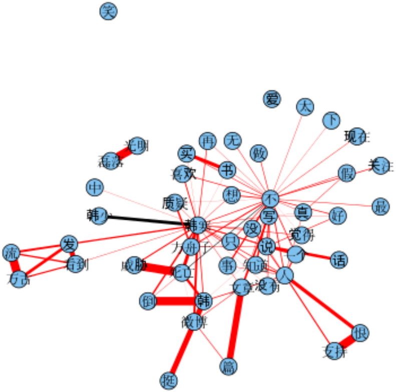
where $||S||_1 = \sum_{i,j}|s_{ij}|$. See, e.g. Banerjee et al. (2007) and Friedman et al. (2007).

Sparse graphical models: results



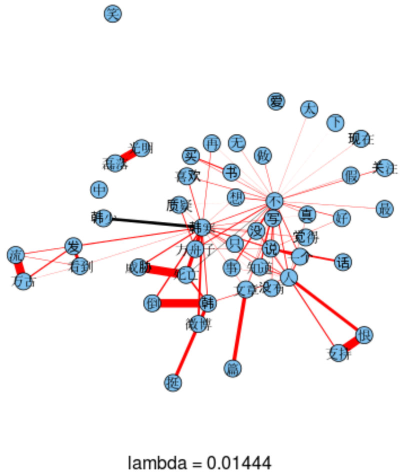
lambda = 0.01

Sparse graphical models: results

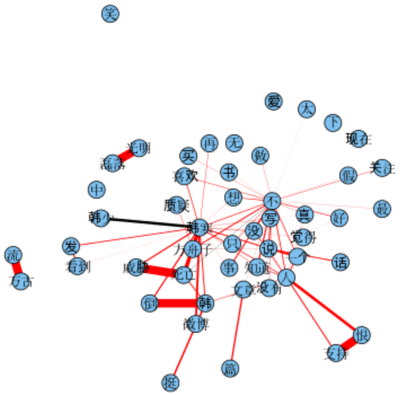


lambda = 0.01222

Sparse graphical models: results

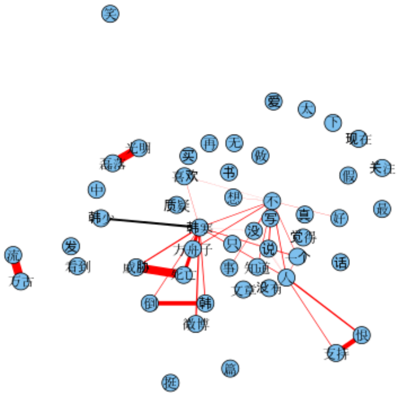


Sparse graphical models: results



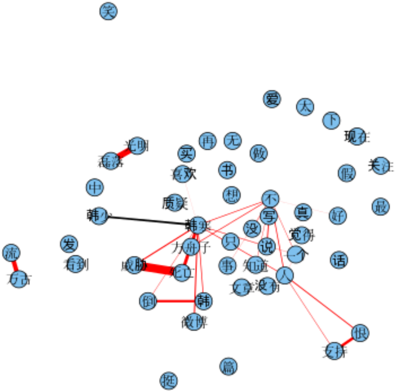
lambda = 0.01667

Sparse graphical models: results



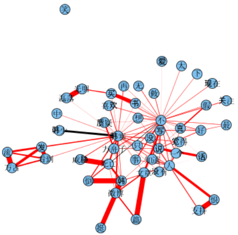
lambda = 0.02556

Sparse graphical models: results

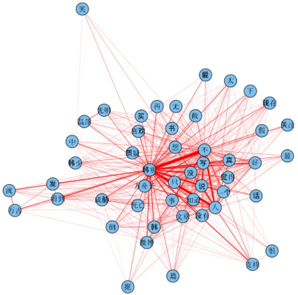


lambda = 0.03

Sparse graphical models v.s. co-occurence



lambda = 0.01



Classification

- $x_i \in R^p$ be the i-th row of $X \in R^{3000 \times 795}$
- y_i the corresponding category. Assume $y_i \in \{-1, +1\}$, where the $+1$ can have the following meanings:
  - positive opinion towards Han Han;
  - negative opinion towards Han Han;
  - netural or unidentifiable opinion;
  - spam.
- Two classification methods: LASSO and l_1-norm SVM.

LASSO

- The Lasso approach (Tibshirani, (1996)):

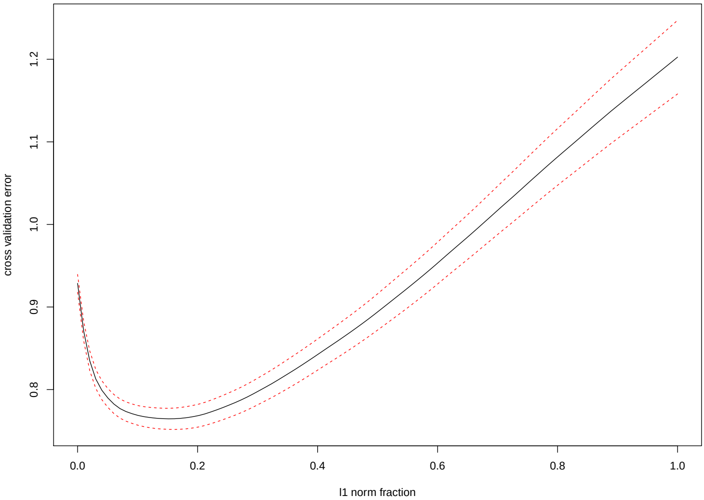
$$\hat{\beta}(\lambda) = \arg\min_{\beta} \frac{1}{2}||y - (\beta_0 + X\beta)||_2^2 + \lambda||\beta||_1$$

- The classifier:

$$\text{class}(x) = \textbf{sign}(\beta_0 + x^T\beta) \in \{-1, +1\}$$

- 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

Choosing λ: cross-validation



LASSO Results

- Three different ways to look at coefficients
- Why: can look at the classifier:

$$\text{class}(x) = \textbf{sign}(\beta_0 + x^T\beta) \in \{-1, +1\}$$

  - Absolute value: most relevant/predictive words
  - Positive: more likely to classify the post in $+1$ category (all other covariates being fixed)
  - Negative: less likely to be in -1 category

LASSO

Positive v.s. Nonpositive classification result

- +1: positive opinion;
- -1: non-positive opinion, including negative, neutral and spam.

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 word images for the positive v.s. nonpositive classification.

Negative v.s. Nonnegative classification result

- +1: negative opinion;
- -1: non-negative opinion, including positive, neutral and spam.

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

LASSO word images for the negative v.s. negative classification.

Standard support vector machine

- Again, linear decision function $f(x) = \beta_0 + \beta x$;
- The classifier $Class(x) = \mathbf{sign}(f(x))$.
- The support vector machine (SVM) (see, e.g. Hastie et al 2001):

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

where $z_+ = \max(0, z)$.

l_1-Norm Support Vector Machine

- Replacing the l_2-norm by l_1-norm yields the sparse SVM (Zhu et al 2003):

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

Positive v.s. Nonpositive classification result

- $+1$: positive opinion;
- -1: non-positive opinion, including negative, neutral and spam.
- Cross validation result:
  - training sample misclassification rate: 16.9%
  - testing sample misclassification rate: 28.2%

Negative v.s. Nonnegative classification result

- $+1$: negative opinion;
- -1: non-negative opinion, including positive, neutral and spam.
- Cross validation result:
  - training sample misclassification rate: 6.4%
  - testing sample misclassification rate: 11.5%

LASSO v.s. l_1-norm SVM

Further Work

- .

- .

- .

- .

- .

- .

- .

- .