文本数据挖掘与分析：初级与高级

孟天广
数据治理研究中心
清华大学政治学系，清华大学苏世民书院

2016年11月25日

内容提要

- 社会科学中的文本分析
- 获取文本数据
- 文本数据分析：基本应用
- 文本数据分析：高级应用
- 自动文本分析举例
- 软件练习

社会科学中的文本分析

- 社会科学对文本展开研究历史悠久，然而，至今文本仍然不是社会科学研究的主流；
- 原因在于：
  - 文本资料难获取；
  - 花时间；
  - 难推广；
  - 难管理；
  - 难分析

大数据时代的文本数据

- 海量文本：体量大、增速快、模态多样
- 百度每天的用户搜索请求，需要1.7天才能扫描一遍；
- 全世界每秒发送290万封电子邮件，一个人需要5.5年日以继夜才能读完；
- 微信每天新增数据达到500TB，比人类所有书籍存量还多；
- 到2020年，数据总量达40ZB，人均5. 2TB

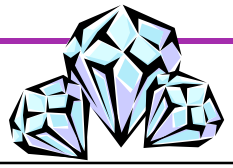
社会科学文本分析的繁荣

- 文本分析繁荣的条件逐步具备：
  - 大规模文本数据采集；
  - 存储和管理能力增强；
  - 文本分析方法蓬勃发展：可推广、系统化和廉价化；
  - 文本资料指数级增长；
  - 通过文本表达的社会意义更广泛；

数据挖掘

- 数据挖掘是知识发现(KDD： Knowledge Discovery in Data)过程中的一个特定步骤，是用专门算法从数据中抽取模式，然后通过解释和评价转换成最终用户可理解的知识
- 数据挖掘是从大量的、不完全的有噪声、模糊的、随机的数据集中识别有效的、新颖的、潜在有用的以及最终可理解的模式的过程。

数据挖掘的过程

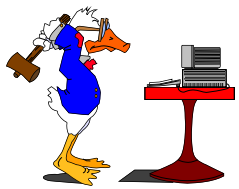


结 果 解 释
和评估

数据挖掘

数据收集
数据预处理

问题定义



文本分析：定义

- 文本分析是收集数据的方法论：研究者采集他人理解世界的信息的途径；
- 由社会结构决定的以语言为载体的社会实践（Fairclough，1989）；
- 行为者表达意义和价值的系统化组织方式（Kress，1993）

文本分析：定义

- 文本分析是研究者描述和阐释一系列记录或可视文本的方法。
- 任何系统化地将文本流降维为一系列可以呈现文本所蕴含的之特征的存在、密度和频次的标准化统计可操作符号的过程（Shapiro and Markoff，1997）；
- 文本分析的目标是描述文本中蕴含之信息的内容、结构和功能；
- 文本分析的任务包括：选择待研究文本的类型、获取文本和选择分析文本的路径。

文本分析：目标

- 表达文本分析(representational)
  - 日常沟通中信息接受者寻求对文本意义尽可能精确地解码；
  - 关注文本的外显(manifest)内容；
- 工具文本分析(instrumental)
  - 关注文本的潜在(latent)内容；
  - 寻找从文本中分析一系列独立于作者原有意图的主题(例如价值偏好、社会心理等)

文本分析：方法路径

- 主题分析(thematic)
  - 识别主题而不是模型化主题间关系；
  - 描绘一系列概念的出现与否；
  - 词频分析
- 语义分析(semantic)
  - 识别主题间的具体关系；
  - 考虑语法、逻辑等；
  - 多种分析结合

大数据与文本分析：机遇

- 通过大规模文本分析探寻新知识
- 文本分析技术的突飞猛进
  - 自然语言过程(NLP);
  - 机器学习(MC);
  - 统计技术引入
- 文本分析软硬件易得！
  - 廉价的计算机;
  - 便捷的互联网;
  - 丰富的开源软件;

习近平总书记讲话（2012-2013）

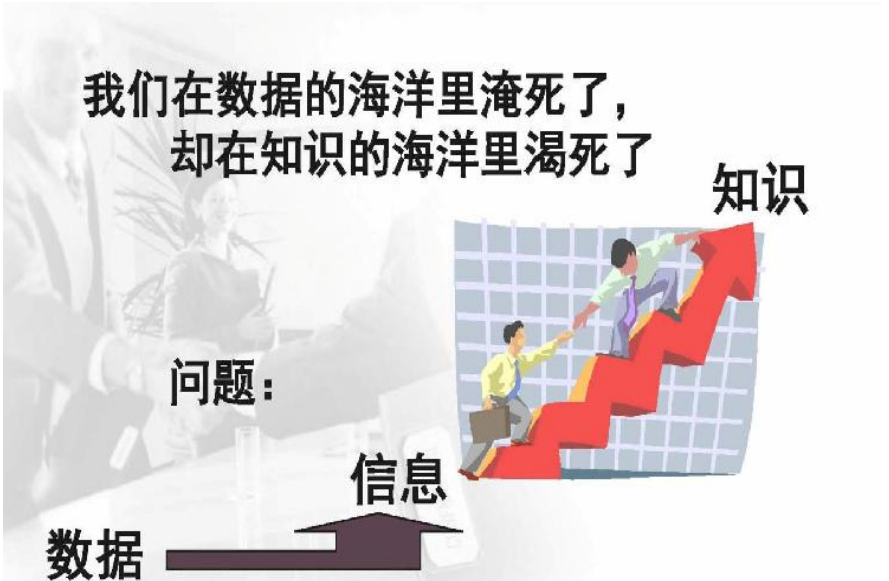
新闻联播中的“国际”与“国内”地域关注

国际地域关注

国内地域关注

大数据与文本分析：挑战

- 非结构化：全世界大约90%的数据以非结构化方式存在
- 海量潜在维度：所有语言的各种可能词语和短语等；
- 文本中词语之间复杂且微妙的关系；
- 词语模糊性和情景敏感性；
- **大数据存在问题：**信息过载、信息失实、 信息冗余、信息污染



- **Big Data in Political Science**
- **Validation: What Big Data Reveal about Survey Misreporting and the Real Electorate**
  Stephen Ansolabehere and Eitan Hersh

- **Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk**
  Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz
- **Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict**
  Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn
- **Reaching Migrants in Survey Research: The Use of the Global Positioning System to Reduce Coverage Bias in China**
  Pierre F. Landry and Mingming Shen
- **An Introduction to Bayesian Inference via Variational Approximations**
  Justin Grimmer
- **Inferential Network Analysis with Exponential Graph Models**
  Skyler J. Cranmer and Bruce A. Desmarais
- **Finding Jumps in Otherwise Smooth Curves: Identifying Critical Events in Political Processes**
  Marc T. Ratkovic and Kevin H. Eng
  Political Analysis (2010) **18(1):** 55-77
- **Improving Predictions Using Ensemble Bayesian Model Averaging**
  Jacob M. Montgomery, Florian M. Hollenbach, and Michael D. Ward
  Political Analysis (2012) **20(3):** 271-291
- **Bayesian Metric Multidimensional Scaling**
  Ryan Bakker and Keith T. Poole
- **The Genealogy of Law**
  Tom S. Clark and Benjamin E. Lauderdale

King etc. (2013). How Censorship in China Allows Government Criticism but Silences Collective Expression

- 主流理论：审查的目标是阻止/打击批评政府
- 作者发现：
- 中国允许新媒体上存在大量批评政府的意见；
- 批评政府意见之所以存在，其原因不在于审查不完美或遗漏，而在于政府审查制度的目标是阻止集体行动，切割社会联系。

数据库

- Collect 3,674,698 social media posts in 85 topic areas over 6 months
- Random sample: 127,283
- For each post (on a timeline in one of 85 content areas):
- Download content the instant it appears
- Revisit each later to determine if it was censored
- Use computer-assisted methods of text analysis

crimson hexagon

KNOW MORE. KNOW WHY. KNOW HOW.

汉网社区 bbs.cnhan.com
读者论坛
便民网 便民论坛 BBS.FZBM.COM
中国红盾交流网
eNet 社区
法治论坛 公平・民主・自由・进
bbs.chinacourt.org
京华论坛
莲池论坛
淮水安澜 bbs.huainet.com
SRT
鲁中论坛 bbs.lznews.cn
雅昌艺术网 论坛 BBS.ARTRON.NET
新版
YNET.com 北青网 青年论坛
ytbbs 烟台论坛
慧聪网 HC360.com
钦州360 .com www.qinzhou360.com
大河论坛 http://bbs.dahe.cn 关注民生 影响中原
金库网 www.JinKu.com
红网 REDNET.CN
外汇论坛
华南论坛
通吃美食 tooeat.com
华声在线 www.voc.com.cn
V1.CN 第一视频新闻网 V1 NEWS
四月青年 社区 bbs.m4.cn
深圳论坛 zbbs.sznews.com
HONGYAN 红岩论坛
CNTV 中国网络电视台
天涯社区 www.tianya.cn
复兴论坛
东湖社区
新街口 社区
洛阳信息港
蓝色沸点
温网论坛 bbs.66wz.com
春秋战国
福建论坛 bbs.66163.com
长城论坛 bbs.hebei.com.cn
文北网京 Beijingww.com
大碗茶论坛
Baidu空间
鳳凰網 論壇 bbs.ifeng.com
55168!
直播吧 www.zhibo8.com
美言网 分享每一天 温暖多一点
江汉热线
合肥热线 www.hefei.cc
光明网 光明社区
站长论坛 .COM
cnwest.com 陕西论坛
新丝路花语
厦门小鱼
中国青年网 youth.cn 青年论坛
名城社区 bbs.2500sz.com
時尚论坛 BBS.TRENDS.COM.CN
搜奇娱乐网 www.qd315.net
sina新浪博客
中原论坛
bbs.macd.cn MACD股票论坛
两江论坛 bbs.cqnews.net
ChinaUnix .net
3e帝国
HONGYAN 红岩论坛
两全其美网校城 中国网校大本营
phpwind
德迷联盟
通吃岛论坛 http://bbs.haogu8.com
全景社区 bbs.p5w.net
CNFOL.COM 中金在线
天互论坛
财经论坛
温陵社区 bbs.qzwb.com
泉州人
浆糊论坛
电脑爱好者俱乐部 CFANCLUB.NET
大金华论坛 分享・消费・信息
315 今天你维权了吗?


设为首页 收藏本站
切换到宽版
金羊社区 bbs.ycwb.com
用QQ帐号登录
只需一步，快速开始
用户名
自动登录
找回密码
密码
登录
注册
首页
论坛
博客
家园
排行榜
快捷导航
请输入搜索内容
帖子
搜索
热搜: 谷俊山案 经适房 征地 狼爸教育 国土局 珠江啤酒 拆迁
抱歉，指定的主题不存在或已被删除或正在被审核
[点击这里返回上一页]
关闭
Sorry, the host you were looking for does not exist, has been deleted, or is being investigated

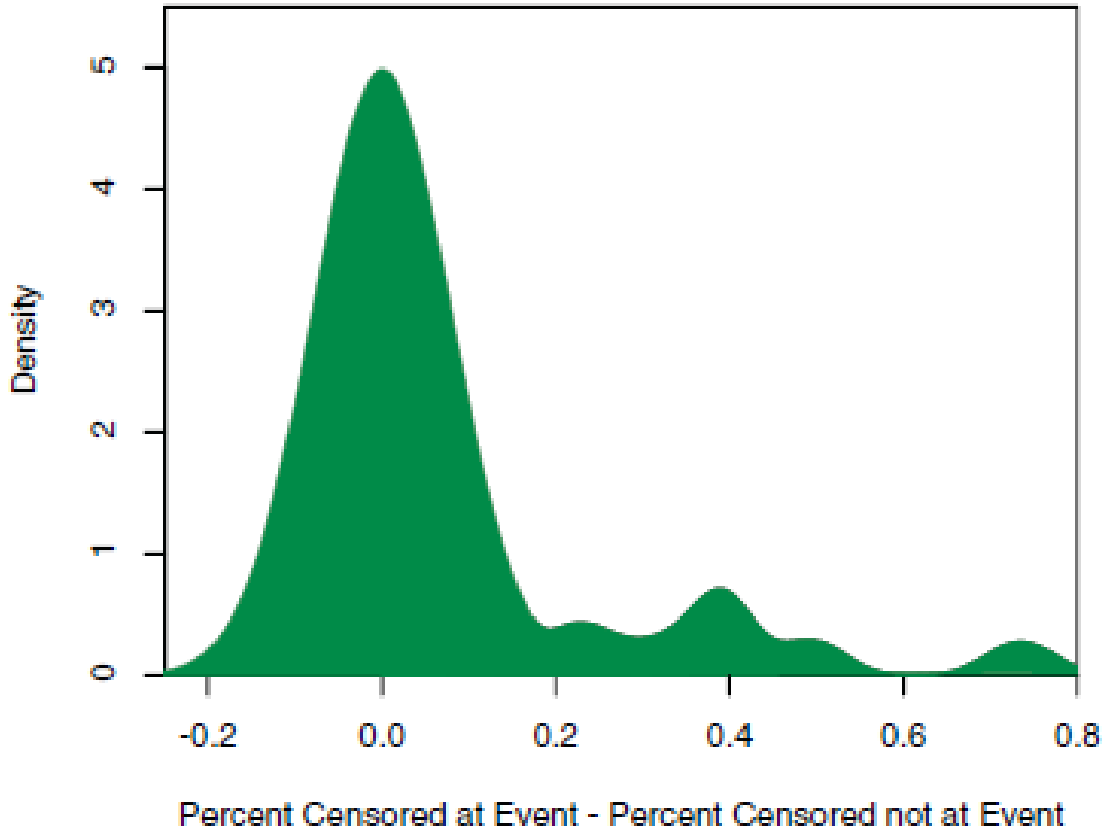
研究假设与分析策略

- 检测85个议题领域；
- 识别发帖量爆发（共发现了87个发帖量爆发），找到相关联的真实事件；
- 选择可能或产生了集体行动的事件；
- 审查与这些事件或发帖爆发相关的所有发帖；
- 判断与这些事件相关的具体帖子，是否与潜在的集体行动相关，或是批评政府
- **政府审查掉所有与潜在集体行动相关之帖子。**

研究设计

- 通过三方面分析来识别因果关系：
  - 发帖量；
  - 与发帖爆发相关之事件是否产生集体行动；
  - 帖子内容；

发帖量

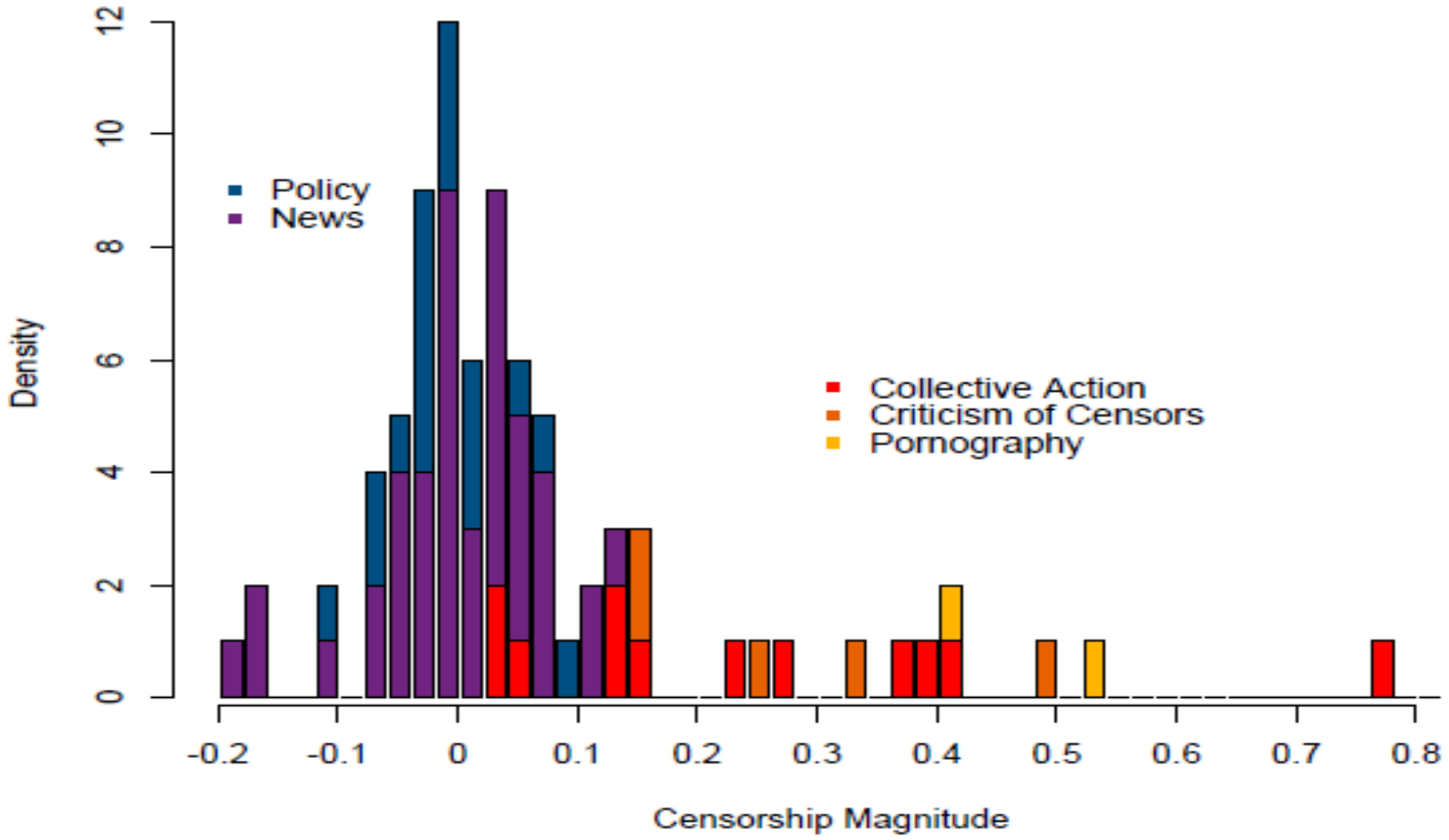


(a) Distribution of Censorship Magnitude

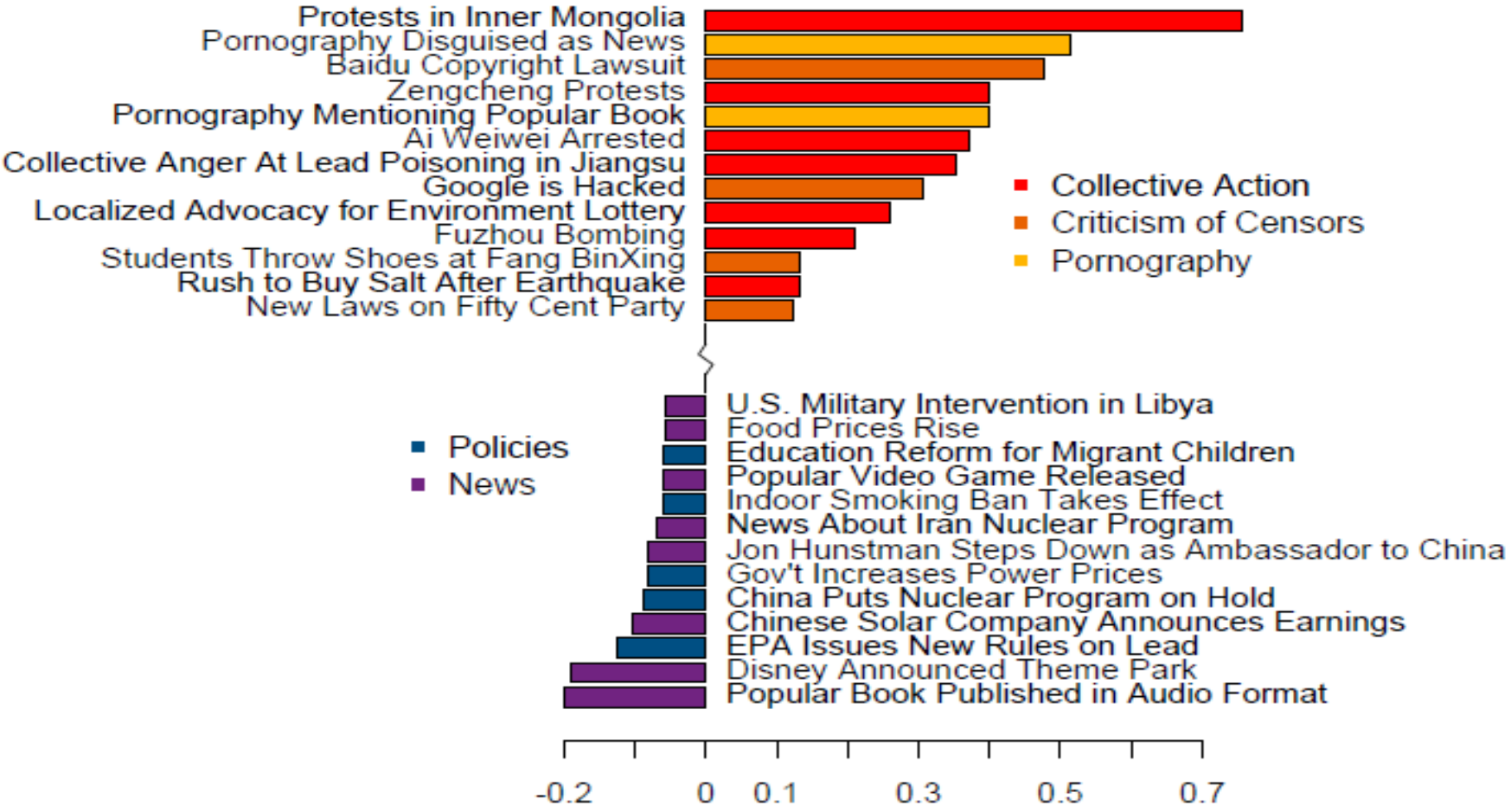
检验二：与发帖爆发相关之事件是否集体行动

- 事件分类：对政府的负面、正面、或中立意见
- 潜在集体行动：
  - 网络外抗议或有组织的群体性事件；
  - 以往有组织或煽动集体行动的个体；
  - 以往煽动抗议或集体行动的民族主义或民族情绪；
- 对审查者的批评
- 色情
- 新闻
- 政府政策

哪类事件被审查？


Density
0
2
4
6
8
10
12
Policy
News
Collective Action
Criticism of Censors
Pornography
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Censorship Magnitude

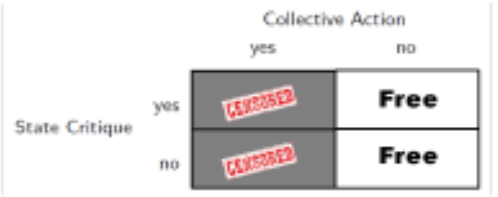
哪类事件被审查？

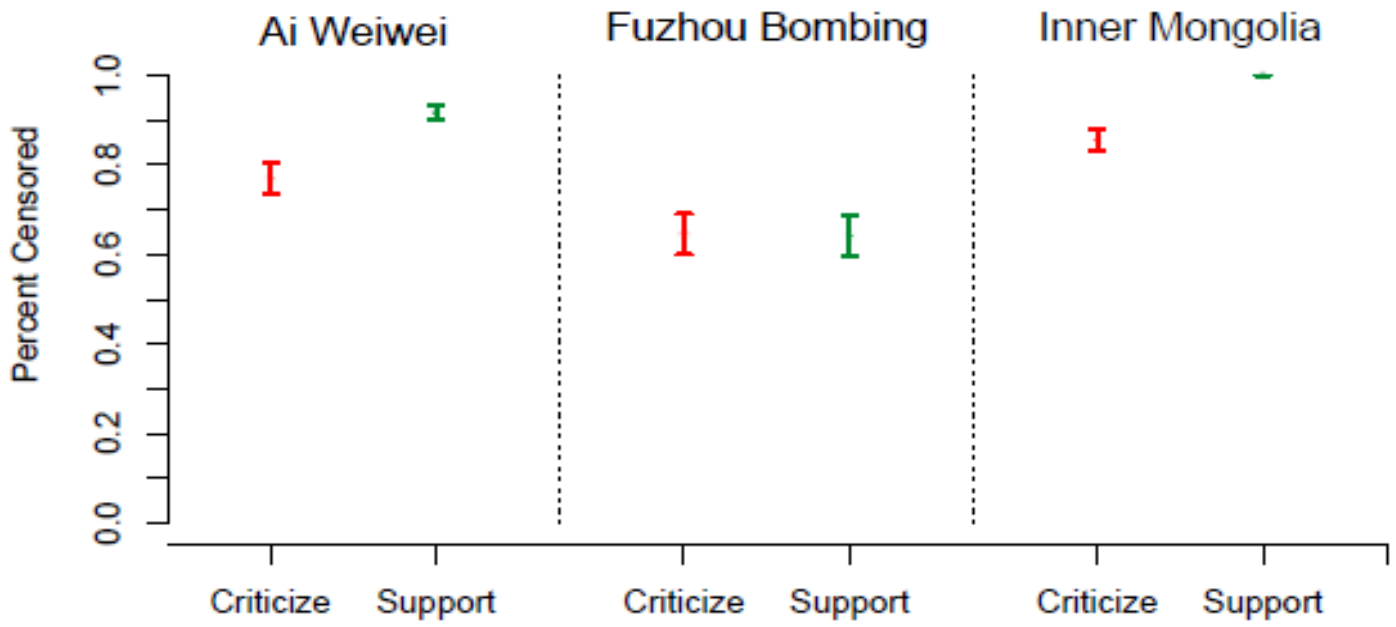

Protests in Inner Mongolia
Pornography Disguised as News
Baidu Copyright Lawsuit
Zengcheng Protests
Pornography Mentioning Popular Book
Ai Weiwei Arrested
Collective Anger At Lead Poisoning in Jiangsu
Google is Hacked
Localized Advocacy for Environment Lottery
Fuzhou Bombing
Students Throw Shoes at Fang BinXing
Rush to Buy Salt After Earthquake
New Laws on Fifty Cent Party
Collective Action
Criticism of Censors
Pornography
Policies
News
U.S. Military Intervention in Libya
Food Prices Rise
Education Reform for Migrant Children
Popular Video Game Released
Indoor Smoking Ban Takes Effect
News About Iran Nuclear Program
Jon Hunstman Steps Down as Ambassador to China
Gov't Increases Power Prices
China Puts Nuclear Program on Hold
Chinese Solar Company Announces Earnings
EPA Issues New Rules on Lead
Disney Announced Theme Park
Popular Book Published in Audio Format
-0.2
0
0.1
0.3
0.5
0.7

检验三：发帖内容

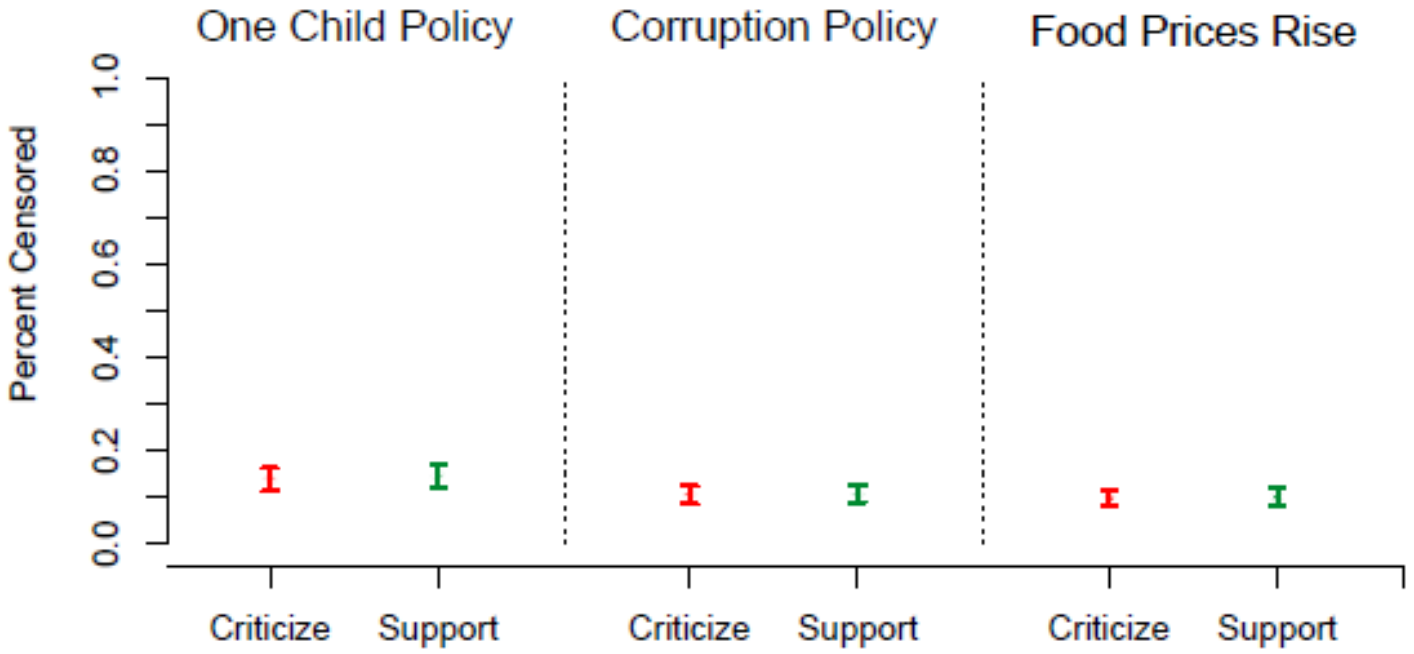
- 采用 “ReadMe” 方法来分析被审查帖子的内容
  - 对中文帖子进行标准化编码；
  - 删除标点符号等；
  - 对字符分段形成词语；
  - 按类别计算被审查掉的词语比例；

With Collective Action, Posts that Support or Oppose the State are Censored





Without Collective Action, Posts that Support or Oppose the State are NOT Censored

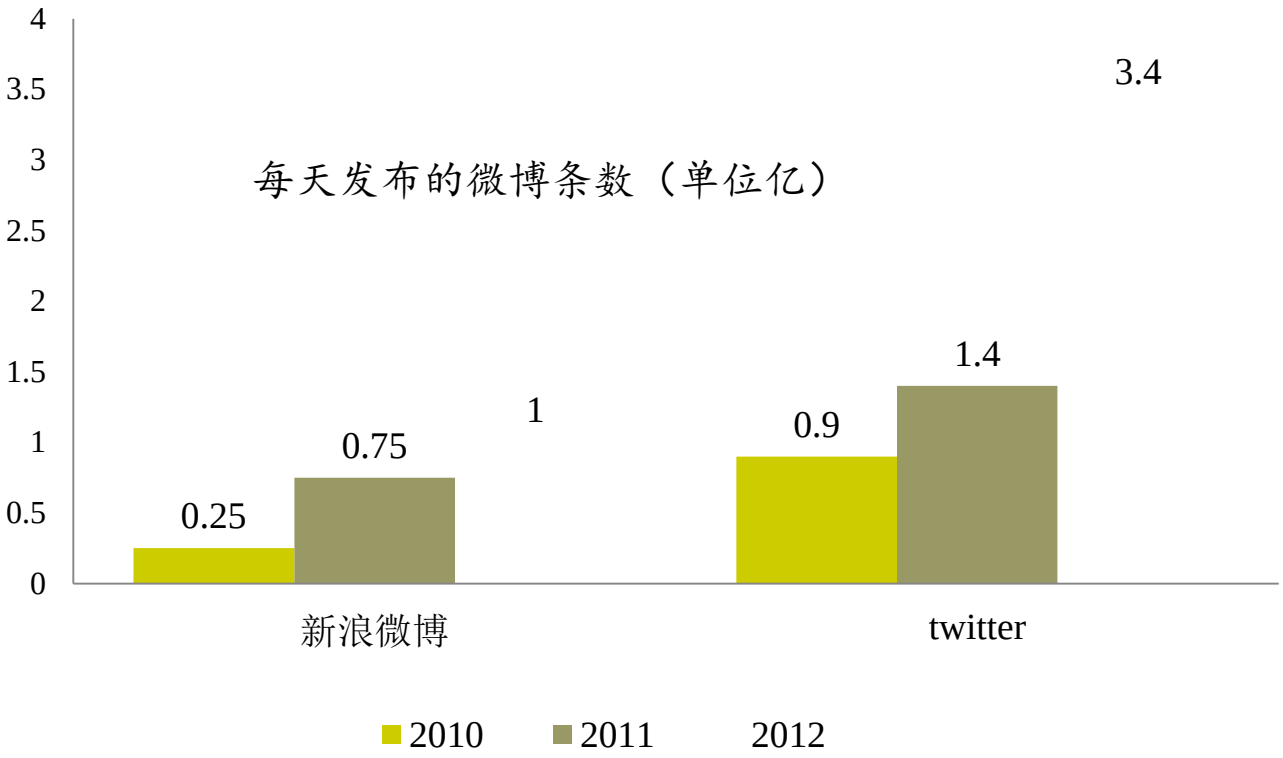


获取文本资料（一）

- 原生数字文本
  - Email/短信
  - 网站HTML
  - RSS feeds
  - 网络社交媒体：微博、Twitter、Facebook
  - 网络论坛：天涯、凯迪社区
  - 网络问答平台：百度百科、Wikipedia
  - 媒体数据库：人民日报、New York Times API
  - 网络交易行为：淘宝、ebay
  - ……

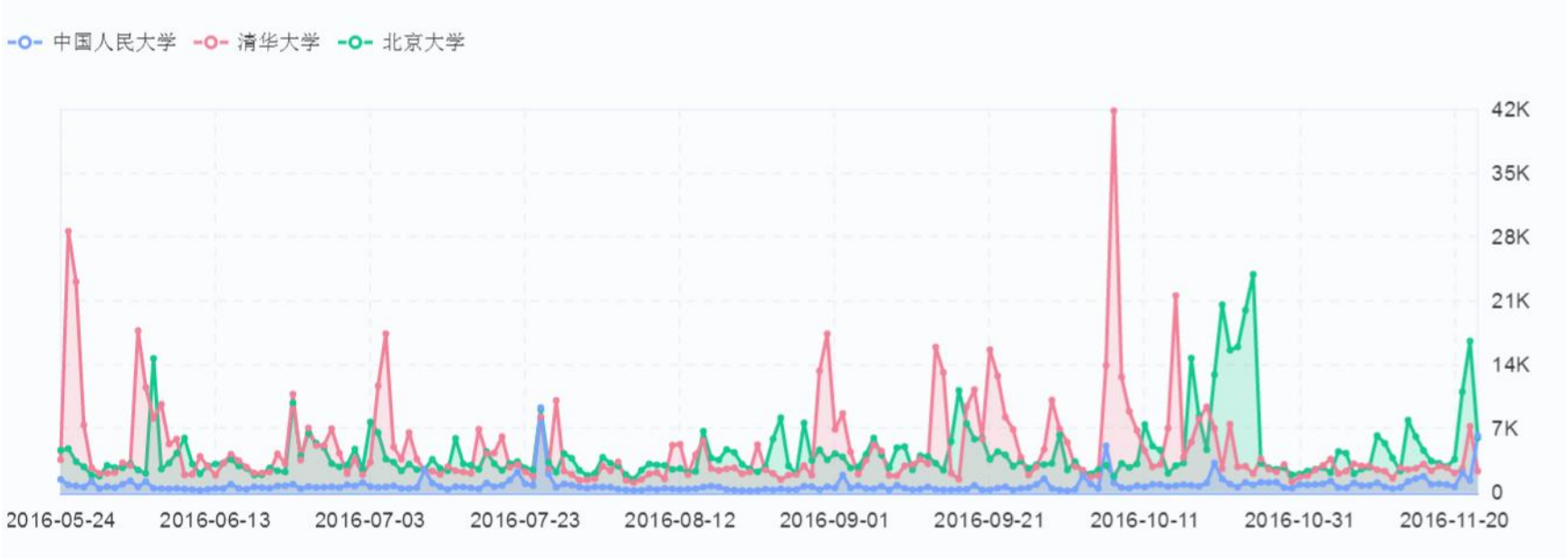
网络社交媒体数据

- 网络社会媒体是目前最大BIG DATA



- Twitter用户4亿，新浪微博1.4亿，微信用户5.5亿
- 微信每日新增数据500TB，QQ每日新增数据200TB

微博数据


中国人民大学
清华大学
北京大学
42K
35K
28K
21K
14K
7K
0
2016-05-24
2016-06-13
2016-07-03
2016-07-23
2016-08-12
2016-09-01
2016-09-21
2016-10-11
2016-10-31
2016-11-20

腾讯大数据

规模数据

QQ月活跃用户数超过	8.4亿
QQ最高同时在线超过	2.0亿
QQ空间月活跃用户数超过	6.5亿
QQ好友关系链对超过	900亿

活跃数据

一天QQ总发消息155亿条，其中发群消息25亿条

一天发空间说说6.5千万条

数据处理

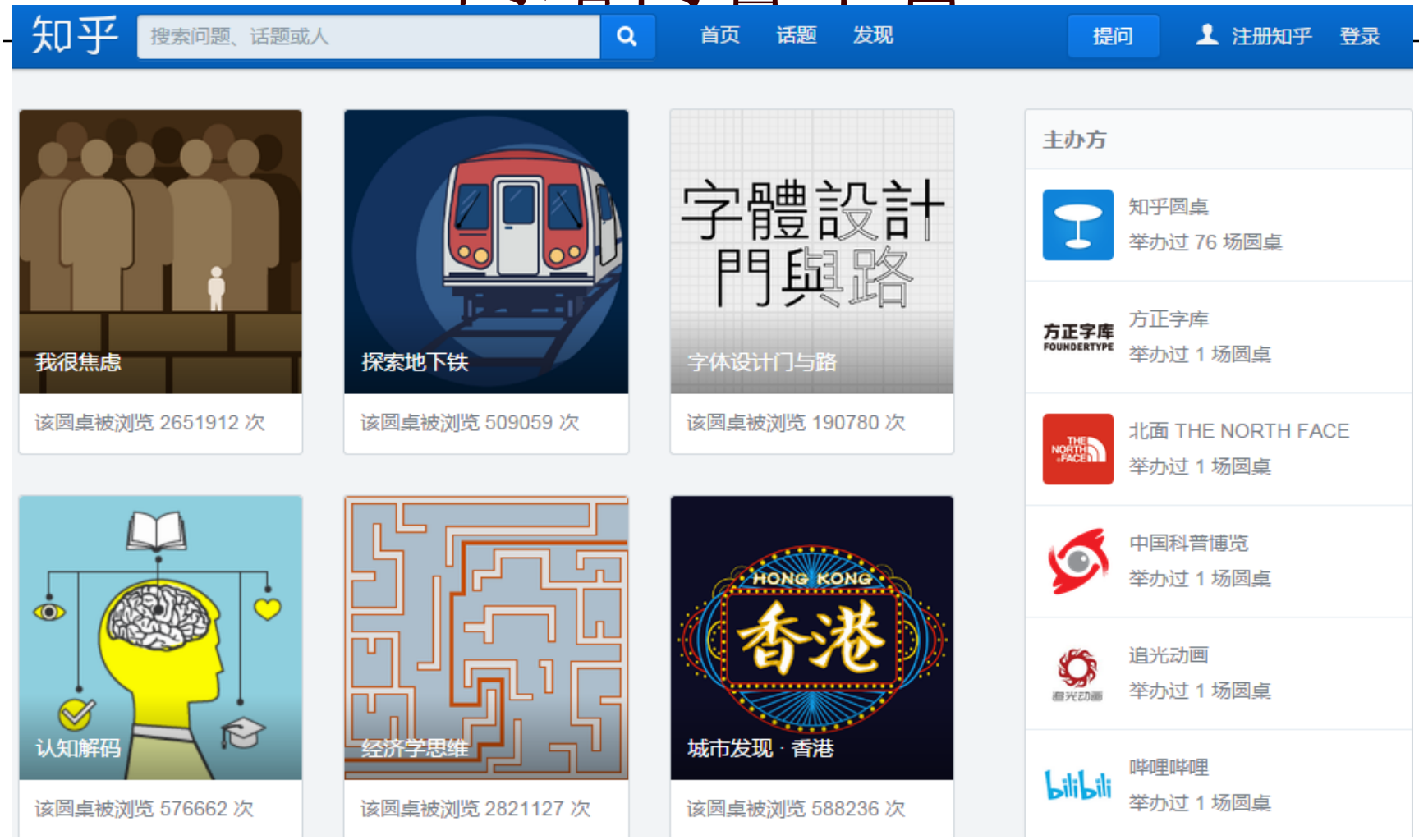
一天接入数据记录1.0万亿条， 新增存储200TB

公共开放数据


中国裁判文书网
Judicial Opinions of China
依法 及时 规范 真实
LAW
首页
民事案件
刑事案件
行政案件
知识产权
赔偿案件
执行案件
文书检索
高级检索>>
关键词:
案号:
裁判时间: 从 2013-7-1
至 2014-1-2
检索说明
搜 索
公报案例
更多>>
各级法院
最高人民法院
北京 天津 河北 山西 内蒙古 辽宁 吉林 黑龙江
上海 江苏 浙江 安徽 福建 江西 山东 河南
湖北 湖南 广东 广西 海南 重庆 四川 贵州
云南 西藏 陕西 甘肃 青海 宁夏 新疆 兵团
最新文书
更
福建省三明市梅列... 冷建军犯危险驾驶罪一审刑事判决书 (2013)梅刑初字第233号 2014-01-02
厦门海事法院 蒋永正、陈锦荣与林增产船舶共有纠纷一审民事判决书 (2013)厦海法商初字第... 2014-01-02
福建省福州市中级... 长乐东港公司与福州市人社局、夏春华社会保障行政... (2013)榕行终字第346号 2014-01-02
福建省浦城县人民法院 马招生盗窃罪一审判决书 (2013)浦刑初字第173号 2014-01-02


政府工作报告

网络问答平台


知乎
搜索问题、话题或人
首页
话题
发现
提问
注册知乎
登录
我很焦虑
该圆桌被浏览 2651912 次
探索地下铁
该圆桌被浏览 509059 次
字體設計
門與路
字体设计门与路
该圆桌被浏览 190780 次
认知解码
该圆桌被浏览 576662 次
经济学思维
该圆桌被浏览 2821127 次
HONG KONG
香港
城市发现 · 香港
该圆桌被浏览 588236 次
主办方
知乎圆桌
举办过 76 场圆桌
方正字库
FOUNDERTYPE
方正字库
举办过 1 场圆桌
北面 THE NORTH FACE
举办过 1 场圆桌
中国科普博览
举办过 1 场圆桌
追光动画
举办过 1 场圆桌
哔哩哔哩
举办过 1 场圆桌

媒体数据库


人民日报
RENMIN RIBAO
10
落实结构性减税 依法加强征管
2010年全国税收收入77390亿元
国内增值税、国内消费税、营业税分别增长14.8%、27.5%、23.8%，所得税增长20.4%
农民变市民 生活怎么样
——湘鄂粤渝四省市城镇化调查
人民观察
为党分忧 为国尽责 为民奉献
——写在第十七届中央纪委第六次全会召开之际
结束德国之行开始对英国进行正式访问
李克强会见英国苏格兰首席部长和苏格兰事务大臣
海南全面推进生态文明建设
春运售票第一天
雷洁琼同志逝世


日报
周报
杂志
人民网 people.cn
人民日报 2011年01月10日 星期一
往期回顾
人民网检索
数字报用户中心
< 上一期
下一期 >
人民日报图文数据库（1946-2014）
01版:要闻
2010年全国税收收入77390亿元
农民变市民 生活怎么样（人民观察）
为党分忧 为国尽责 为民奉献
李克强会见英国苏格兰首席部长和苏格兰事务...
海南全面推进生态文明建设
雷洁琼同志逝世
春运售票第一天
版面目录
第01版：要闻
第02版：要闻
第03版：要闻
第04版：要闻
第05版：要闻
第06版：要闻
第07版：理论
第08版：视点
第09版：广告
第10版：经济

获取文本资料（二）

- 数字化档案
  - CNKI等网络档案库
  - 新浪爱问等资料分享平台
  - Google Books、百度学术
  - JSTOR Data for Research
  - 年鉴数据库
  - 报刊数据库
  - ……


信息科技
经济与管理科学
报纸名称
中国计算机报 (190)
人民邮电 (170)
计算机世界 (127)
网络世界 (104)
科技日报 (103)
研究层次
行业指导(社科) (1957)
工程技术(自科) (349)
职业指导(社科) (265)
政策研究(社科) (245)
大众文化 (45)
检索历史：
大数据
最低生活
分组浏览: 学科 发表年度 作者 单位
免费订阅 定制检索式
2014(1374) 2013(1169) 2012(323) 2011(55) 2010(12) 2009(6) 2008(5) 2007(5) 2006(7) 2005(4)
2004(3) 2003(2) 2002(2) 2001(1)
排序: 主题排序 报纸日期
切换到摘要 每页显示: 10 20 50
(0) 清除 导出 / 参考文献 分析 / 阅读
找到 3,108 条结果 浏览1/63 下一页
题名 作者 报纸名称 日期 下载 预览 分享
1 我"对地观测大数据应对全球变化"获联合国奖项 李大庆 科技日报 2014-09-04
2 创新旅游大数据 陈罡 科技日报 2014-09-04
3 亿万数据自动检索 目标车辆瞬间锁定 孙丽丽 通讯员 董月亮 刘冰 人民公安报 2014-09-03
4 大数据能有效解决市民出行拥堵难题 安吉 科技日报 2014-09-03
5 临沂：以大数据云计算为支撑提升打防管控能力 张黎明 人民公安报 2014-09-02
6 大数据的价值 杨洪涛 光明日报 2014-09-01
7 有了大数据，农业种植不看天 贾婧 科技日报 2014-08-30
8 为大数据产业发展提供全方位服务 刘志强 科技日报 2014-08-30

Google 大数据

网页 图片 新闻 视频 图书 更多 搜索工具

找到约 1,570,000 条结果 （用时 0.62 秒）

数据挖掘原理与算法 - 第 78 页

books.google.com/books?isbn=7508416538

邵峰晶于忠清 - 2003 - 部分预览 - 更多版本

1 数据挖掘不同领域中的采样数据挖掘中的采样方法很多,我们只足限制在常用的数据挖掘任务:关联规则、分类、聚类及这些算法在大数据集时的应用。 1 . 关联规则的采样挖掘关联规则的任务通常与事务处理与关系数据库相关,该任务需要反复遍历 ...

大数据时代

books.google.com/books?isbn=7213052543 - 转为简体网页

Viktor Mayer-Schönberger, 邁爾, Kenneth Cukier - 2013 - 无预览

软件测试方法和技术 - 第 165 页

books.google.com/books?isbn=7302111332

2005 - 部分预览 - 更多版本

Google Ngram

between 1900 and 2000 from the corpus Chinese (simplified) with smoothing of 3 . Search lots of books

获取文本资料（三）

- 自助文本资料扫描+OCR
- 图书馆资料；
- 访谈记录；
- 历史档案；
- 照片、图片等；
- 回忆录；
- ……

中国历代人物数据库

China Biographical Database Project (CBDB)

Hi, Guest login

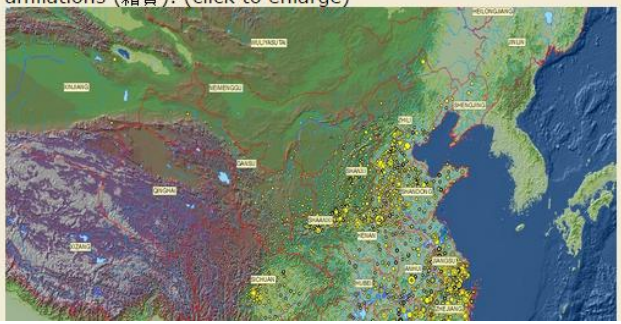
Home | About Us | Funders of CBDB | Sources and Coverage | Methodology | Conferences and Workshops | Download | 中文 | Wilkinson

Home

Welcome

Introduction

The China Biographical Database is a freely accessible relational database with biographical information about approximately **360,000 individuals as of April 2015**, primarily from the 7th through 19th centuries. With both online and offline versions, the data is meant to be useful for statistical, social network, and spatial analysis as well as serving as a kind of biographical reference. The image below shows the **spatial distribution** of a cross dynastic subset of 67,000 people in CBDB by basic affiliations (籍貫). (click to enlarge)



Updates

Tang Research Foundation and CBDB

We are pleased to announce that beginning in January 2015 the Tang Research Foundation is supporting a project for a comprehensive prosopographical investigation of the Tang period based on excavated epitaphs (墓誌), official documents, and private writings.

New Release

UPDATED CBDB database as of April 2015. For download and information about this release follow this link:

http://isites.harvard.edu/icb/icb.do?keyword=k16229&tabgroupid=icb.tabgroup144476

Signing up for the CBDB mailing list

网络文本抓取

- 网络空间存在海量文本数据
- 网络爬虫软件：是一种按照一定的规则，自动的抓取网络信息的程序或者脚本。
- 从网页中抽取文本信息；
- 网页文本清理：
  - 清楚乱码、计算机语言代码
  - 停词、标点符号、大小写
  - 表情、图片、网络链接等

网络文本抓取软件

- Rweibo：R抓取新浪微博的包，基于新浪API进行开发，可以搜索、抓取新浪微博数据；
- XML：R抓取网络页面的包；
- Scrapy：Python开发的一个快速、高层次的屏幕抓取和web抓取框架，用于抓取web站点并从页面中提取结构化和非结构化数据；
- lxml： Python开发的简洁快速抓取网页内容的包；
- 网络数据采集器：GooSeeker；八爪鱼采集器；火车头采集器等

爬虫软件


八爪鱼采集器
最好用的网页数据采集器，让数据触手可及！
免费下载
1分钟快速了解
任何人都可以使用
还在研究网页源代码和抓包工具吗？现在不用了，会上网就能采集，所见即所得的界面,可视化流程，无需懂技术，点点鼠标，2分钟即可快速入门。
任何网站都可以采集
不仅使用简单，而且功能强大：点击，登陆，翻页，甚至识别验证码，当网页出错误，或者多套模版完全不一样的时候，还可以根据不同情况做不同的处理。
云采集，关机也可以
配置好采集任务，就可以关机了，任务可以在云端执行，数量庞大的企业云，24*7不间断运行，再也不用担心IP被封，网络中断了，还能瞬间采集大量数据。


GooSeeker 集搜客
首页
产品
资源
教程
社区
wenq10
定制服务
集搜客GooSeeker网页抓取软件
玩转大数据，发现数据之美
8年打造品牌产品
免费下载
观看介绍视频
了解产品 >
免编程抓取数据
模板资源套用
通用网络爬虫
不限深度和广度

文本数据分析：基本应用

- 分词与停词
- 词频
- 词云
- 词语索引
- 词语搭配
- 文本比较
- 特定词识别（时间、地址、网址等）

文本数据的属性

- ？？
  - 你感兴趣文本数据的哪些属性？
  - 文本数据的哪些属性可以被分析？

文本分析的难与易？

您好！我是居住在仪征市真州镇中桥村娄庄的居民，地处扬州化工园区内，目前我们面临如下一些问题，还请领导能够关心一下：1、村庄附近已建成两家化工厂，其中一家石友化工在2013年和2014年分别发生了1次火灾，这让居住在附近的我们很恐惧，然园区为稳定民心象征性组织了一次拆迁评估，截止目前为止此事已不了了之；2、今年4月就在村庄东侧又开始兴建化工厂，园区多次承诺当有项目来此地征用土地时，一定会将附近居民安置到其他地方，现在化工厂已完工一半，园区代表再次改口，不顾及本地居民的人身安全；3、自2012年起园区就拆迁一事一直拖延至今，不予解决，给我们的生活造成了很多的不便，出行道路遭到破坏，各种农作物种植入不敷出导致现在已无人种植农作物，各种生活垃圾工业垃圾被倾倒在村庄周围。4、空气污染严重以上是我们目前遇到的一些问题，还请领导能在百忙之中抽出时间审阅一下，再次表示诚挚的谢意！

文本分析的难与易？

一	10
在	8
的	7
工	7
园	5
次	5
了	5
区	5
化	5
地	5
不	5
我	5
居	5
园区	5
居民	3
附近	3
村庄	3
居住	2
化工	2
拆迁	2
种植	2
化工厂	2
垃圾	2
领导	2
各种	2
农作物	2
再次	2
问题	2

文本清理

- 分析文本资料之前需要清理文本
  - 从人的表达方式到机器的理解方式
  - 删除停词；
  - 大小写；
  - 提取词干；
  - 去除标点符号；
  - 去除非字符符号；
  - 提取标题、作者、日期、电子邮箱、地址等特殊功能词
  - 等等

分词（tokenizer）

- 分词即确定每个词的边界，把句子划分与一个个词组。
- 对于英文文本，由于每个单词都有空格分隔，英文文本已经天然的完成分词工作。所以分词主要是针对中文文本的。
- 如可以利用词性（名词、动词、形容词，等等）。

分词前

您好！我是居住在仪征市真州镇中桥村娄庄的居民，地处扬州化工园区内，目前我们面临如下一些问题，还请领导能够关心一下：1、村庄附近已建成两家化工厂，其中一家石友化工在2013年和2014年分别发生了1次火灾，这让居住在附近的我们很恐惧，然园区为稳定民心象征性组织了一次拆迁评估，截止目前为止此事已不了了之；2、今年4月就在村庄东侧又开始兴建化工厂，园区多次承诺当有项目来此地征用土地时，一定会将附近居民安置到其他地方，现在化工厂已完工一半，园区代表再次改口，不顾及本地居民的人身安全；3、自2012年起园区就拆迁一事一直拖延至今，不予解决，给我们的生活造成了很多的不便，出行道路遭到破坏，各种农作物种植入不敷出导致现在已无人种植农作物，各种生活垃圾工业垃圾被倾倒在村庄周围。4、空气污染严重以上是我们目前遇到的一些问题，还请领导能在百忙之中抽出时间审阅一下，再次表示诚挚的谢意！

分词后

您好 我 是 居住 在 仪征市 真州镇 中桥村 娄庄的 居民 地处 扬州 化工 园区 内 目前 我们 面临 如下 一些 问题 还 请 领导 能够 关心 一下 1 村庄 附近 已 建成 两 家化 工厂 其中 一家 石友 化工 在 2013 年和 2014 年 分别 发生 了 1 次 火灾 这 让 居住 在 附近 的 我们 很 恐惧 然 园区 为 稳定 民心 象征性 组织 了 一次 拆迁 评估 截止 目前 为止 此事 已 不了了之 2 今年 4 月 就 在 村庄 东侧 又 开始 兴建 化工厂 园区 多 次 承诺 当 有 项目 来 此地 征用 土地 时 一定会 将 附近 居民 安置 到 其他 地方 现在 化工厂 已 完工 一半 园区 代表 再次 改口 不顾 及 本地 居民 的 人身 安全 3 自 2012 年起 园区 就 拆迁 一事 一直 拖延 至今 不予 解决 给 我们 的 生活 造成 了 很多 的 不便 出行 道路 遭到 破坏 各种 农作物 种植 入 不敷出 导致 现在 已 无人 种植 农作物 各种 生活 垃圾 工业 垃圾 被 倾倒 在 村庄 周围 4 空气 污染 严重 以上 是 我们 目前 遇到 的 一些 问题 还 请 领导 能 在 百忙之中 抽出 时间 审阅 一下 再次 表示 诚挚 的 谢意

停词（Stop words)

- 停词是那些不承载信息的的词语；
  - 通常扮演功能性角色；
  - 是文本中不传递信息的噪音；
  - 丢弃停词是为了关注实词（substantive words)
- 英语：a, about, above, across, after, again
- 中文：在，的，也，地，它，是，就，和
- 停词可自定义

中文常用停词表

- Baidu停词表；
- http://blog.csdn.net/shijiebei2009/article/details/39696571；
- 哈工大停用词表；
- 自定义停词表

词云（Word Cloud）

- 词云是对文本中出现频率较高的“关键词”予以视觉上的突出，形成“关键词云层”或“关键词渲染”，从而过滤掉大量的文本信息，使文本主旨信息直接呈现。

词云举例：五中全会公报


发展
建设
创新
经济
坚持
全面
社会
全会
战略
人民
安全
完善
推动
必须
环境
改革
加快
促进
中央
深化
小康
科学
服务
开放
深入
思想
水平
精神
我国
重大
社会主义
现代化
生态
体系
增强
习近平
十八
基础
质量
行动
总书记
同志
全体
计划
保护
和谐
公共
农业
政府
保障
人才
人口
自然
会议
意识
贫困
布局
全国
团结
实力
引领
委员会
就业
领域
实行
开展
特色
培育
规律
中央委员
理论
政策
加大
现代
农村
群众
统筹
市场
文化
工程
各族
贡献
港澳
全党
科技
创业
节约

文本比较

- 比较不同文本间（文档间）一系列文本特征的异同；
- 比较对象：文本长度、用词、词频、关键词等；
- 研究不同政治立场、文化、组织间通过文本表达的主张、价值和意义

其他文本分析方法

- 文本摘要：从文档中抽取关键信息，用简洁的形式对文档内容进行摘要或解释。
  - 用户不需要浏览全文就可以了解文档或文档集合的总体内容；
  - 篇首截取法、上下文截抽取法、论题句抽取法、仿人法等
- 分布分析：通过对文档的分析，得到特定数据在特定历史时刻的分布情况
- 词性分析
- 话题分类

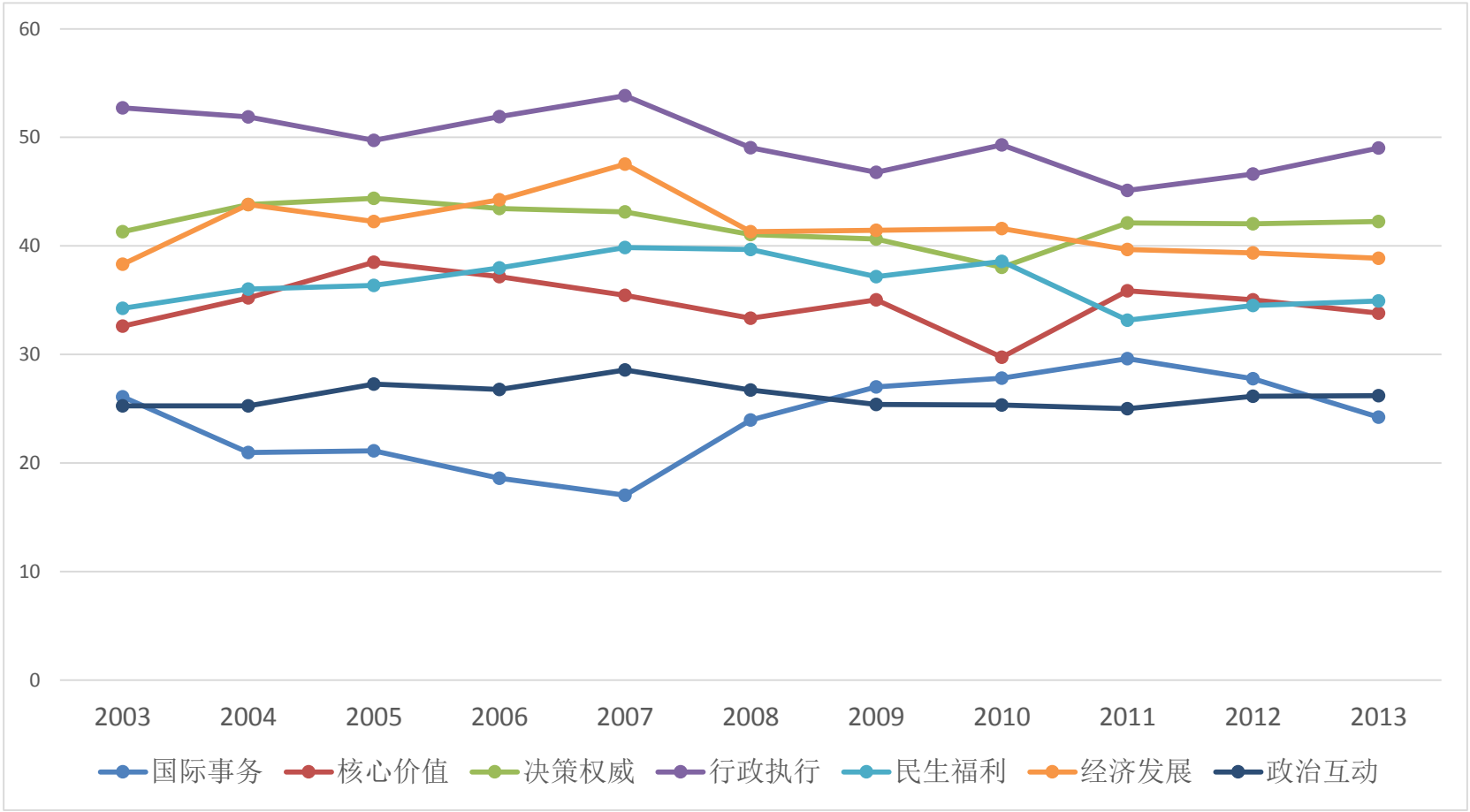
词性：2003-2013《新闻联播》

词性	词汇频率	比例（%）
名词	2,884	59.35
动词	1,422	29.27
形容词	217	4.47
副词	143	2.94
数量词	133	2.74
其他	60	1.23
合计	4,859	100

话题分类

- 话题分类指按照研究者预先定义的话题类别，对文档集合中的每个文档赋予一个特定类别的过程。
  - 目标是了解不同话题在总体中的比例、变化趋势等
  - 第一步，定义话题类别；
  - 第二步，分析每个文章，赋予其特定话题名称；
  - 第三步，分析各话题的分布及趋势等属性；
  - 第四步，精炼话题定义，重复上述过程。

话题分类：2003-2013《新闻联播》



中文文本分析开源软件

- Rwordseg：R环境下的中文分词工具，使用rjava调用java分词工具ansj.
- wordcloud：R环境下词云制作工具
- 中科院ictclas分词工具包；
- FudanNLP是为中文自然语言处理而开发的java工具包
- Jieba：Python中中文分词工具包

文本数据分析：高级应用

- 文本分类
- 主题模型
- 情感分析
- 语义网络分析

文本分类

- 文本分类是指按照预先定义的主题类别，为文档集合中的每个文档确定一个类别。
- 文本分类的目标是识别区别不同类型文本的特征模式
- 如何找到这些特征模式？
  - 词典法；
  - 人工分类：人工识别与不同类型文本相关的词语；
  - 有监督或无监督机器学习的分类方法：依赖于统计模型

文本分类：词典法

- 词典法：基于识别的词语系统来分类。
- 文档清理和分词；
- 定义分类模式；
- 文档，部分根据分类模式标示；
  - 训练集：用于开发词典或利用已有词典；
  - 测试集：用于检验词典；
  - 分类集：使用词典对未标示文档进行分类；
- 根据词语出现频次来分类；
- 分类未标示文档

利用词典法进行文档分类

分类：拆迁

定义词典：

拆迁 补偿 安置 土地 开发商 改造 回迁 动迁 征收 赔偿 征地 拆迁户 搬迁 拆除 棚户区 危房 宅基地 征用

分类：环保

定义词典：

污染 环境 严重 垃圾 环保 噪音 健康 空气 污水 环保局 气味 臭气 扰民 施工 噪声 卫生 焚烧 化工厂 刺鼻 气体 油烟 粉尘

尊敬的省委书记；新年好！我是德兴市茶叶示范场拆迁户，在2012年单位开发工业园区，我1986年建好的房屋没有房产证单位不给上报房管所，说没房产证不给安置，我没办法住在出租房，最近分房了，才发现别人2011年建的房屋很多人没房产证都有安置，说我的房屋是违章建筑，在2012年8月20日被强拆，我们没有签字，单位主管跟我有过节，我的1995年房改的房子单位是分了我，当时我已有自建房，我侄子要我给他结婚住，一年后离婚了，为了收回装修费，在我不知情下卖给人家了，当时我在外打工，单位就说我有房的，卖了，现在不给安置了，我俩老天天为住房求来求去，要求省委书记帮忙我俩解决住房问题！

苏书记您好：我是萍乡市排上镇排上村村民，我们镇上方圆一公里内居然就有两个铅污染严重的厂，还有一个离镇中心也只有三四千米远。其中有个座落在山腰上，镇上的自来水厂就是山的另一侧。铅污染的危害众所周知，请苏书记您能从百忙中处理一下这个问题，我代表排上村数千居民恳求您了

有监督机器学习法

- 有监督学习法的元素：人工分类+机器学习
- 类别集合
  - 信用评价，政治主张等
  - 正面态度-负面态度
  - 亲美国，态度模糊，反对美国
- 手动编码文档集合
  - 人工编码
  - 训练集：手动编码后让机器学习编码规则
  - 检验集：用于检验编码结果好坏
- 未编码文档集合
  - 基于手动编码训练集来编码未标示文档的方法

无监督机器学习法

- 无监督学习法的元素：机器学习
- 定义类别集合的数量
- 分析过程
- 计算机依据词语出现频次或共显关联自动分类所有文档
- 呈现特定类别文章的高频词
- 研究者依据高频词为每个类别命名
- 检验集：检验自动编码结果好坏

两种有监督机器学习法

- Naive Bayes: 关注每个单个文档的分类
- ReadMe: 仅仅关注总体中每个类别的比例

自动文本分析的前景与陷阱

- 三类文本分类（classification）方法（Grimmer，2013）：
  - 字典法（dictionary methods），根据关键词的出现次数来确定；
  - 有监督学习法（supervised learning methods），先由人工构建编码练习库，然后让机器根据人工编码模式来进行自动编码，最后将机器编码与人工编码相比较检验其效度；
  - 无人监督学习法（unsupervised learning methods）不需要人工事先编码，而是基于模型假设和文本性质来分类并自动将文本分配到各类别。

自动文本分析的前景与陷阱

- 自动文本分析确保研究者便捷地实现文本分类和定位，但仍存问题：
  - 机器自动识别有很多不准确的地方，但仍然在很多方面给学者提供了研究便利；
  - 自动文本分析不能取代学者的阅读和思考；
  - 没有一个最完美的自动识别方法；
  - 对自动文本分析结果的效度分析非常重要。

无监督机器学习法

- 聚类分析(Cluster analysis)；
- 主题模型(Topic modeling)；

聚类分析（cluster analysis)

- 文本聚类是将文档集合分成若干个簇，要求同一簇内文档内容的相似度尽可能地大，而不同簇间的相似度尽可能地小。
- K-means clustering
  - 文档库；
  - 测量距离；
  - 定义K；
- 对于K群中任意一群以任意一随机文档为中心；
- 每个文档被归类到距离最近的文档群；

距离测度(Distance Metrics)

- 欧氏距离(Euclidian distance):

$$L_2(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$

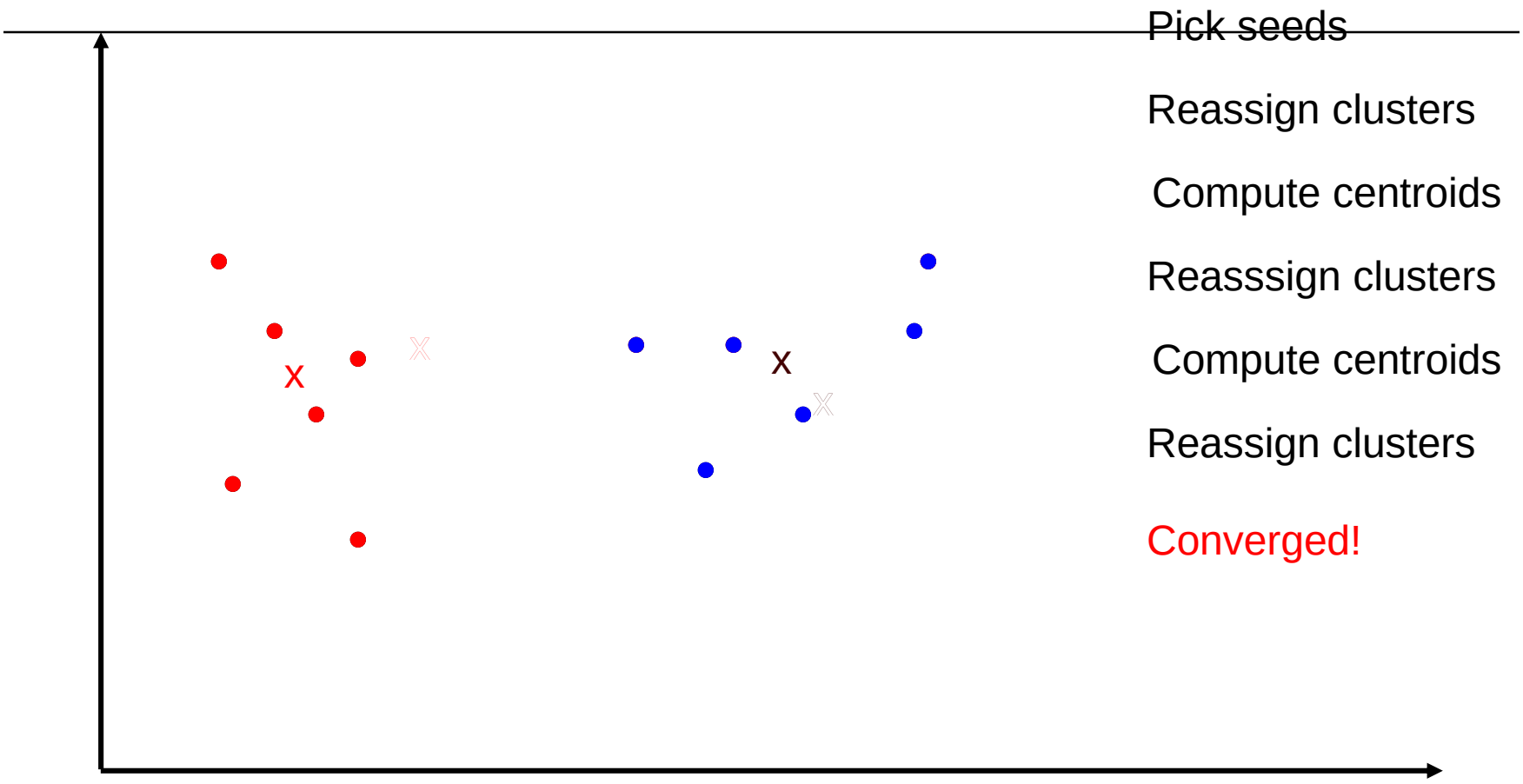
- 曼哈顿距离(L_1 norm):

$$L_1(\vec{x},\vec{y}) = \sum_{i=1}^{m}|x_i - y_i|$$

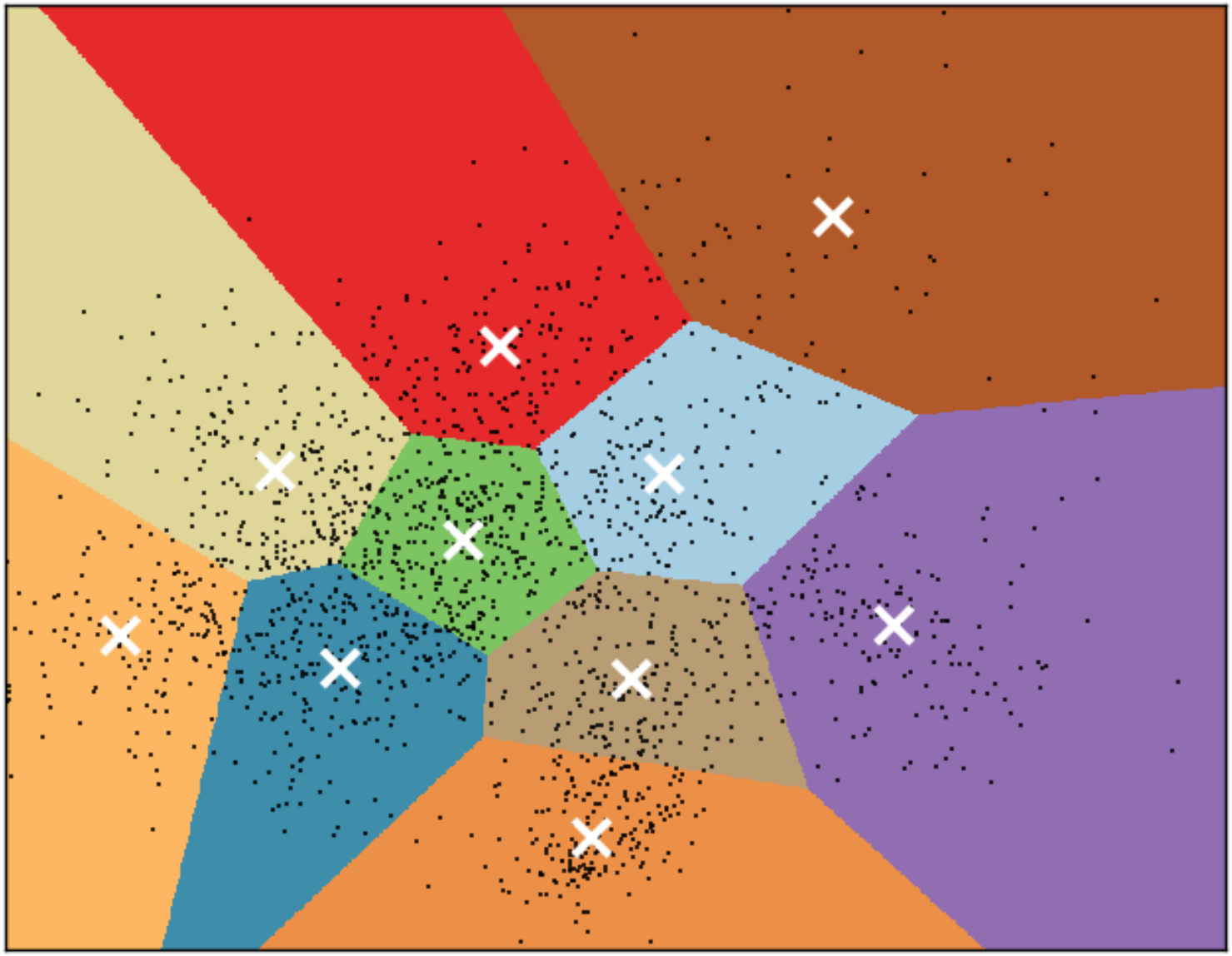
- 余弦相似度(Cosine Similarity):

$$1 - \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \cdot |\vec{y}|}$$

K-Means Example(K=2)



K-means clustering on the digits dataset (PCA-reduced data)
Centroids are marked with white cross



主题模型

- Topic model 是一种应用十分广泛的产生式模型(generative model)，是一系列将统计学、文本分析和机器学习相结合的文本分类方法。
  - 语料库存在一系列潜在主题
  - 由词语出现概率与词语间共现关系来推论潜在主题
  - 由潜在主题来呈现代表性词语
  - 类似于因子分析的降维

主题模型的基本假定

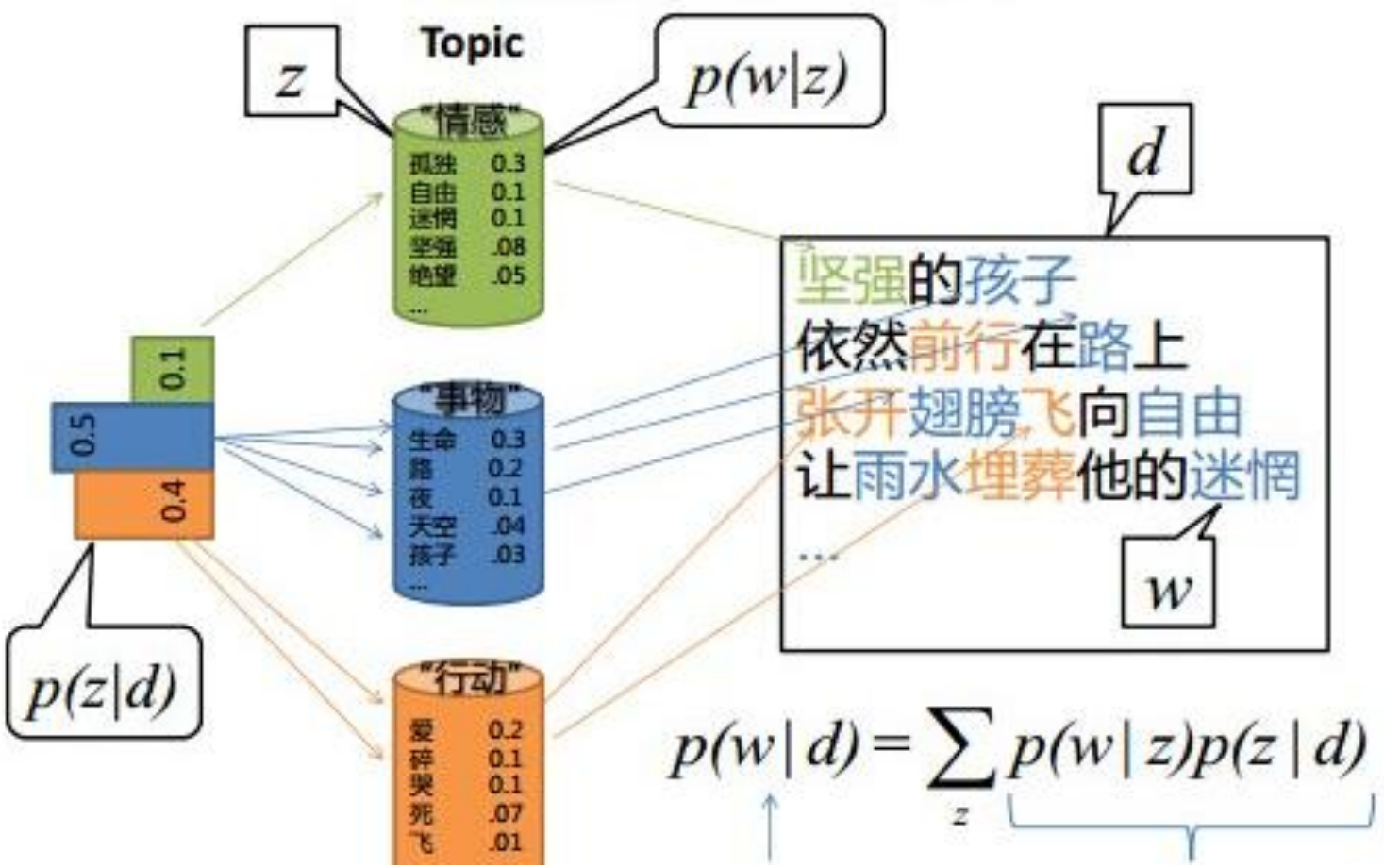
- 假定1：词语与主题相关；
- 假定2：不同词语集合与不同主题相关；
- 利用词语出现概率来呈现主题；
- 主题实际上是词语及其集合的概率方程

LDA

- LDA（Blei et al., 2003）：主题模型的基本形式。
- LDA认为一个离散数据集合（如文档集合）是由隐含在数据集合背后的topic set 生成的，这个set中的每一个topic都是词的概率分布。
- 利用已知的文档-单词信息$P(w|d)$，训练出单词-主题$P(w|z)$，和文档-主题$P(z|d)$，挖掘文本中潜在的语义知识，实现文档和单词的主题聚类

$$P(w|d) = P(w|z) * P(z|d)$$

LDA举例


Topic
z
p(w|z)
"情感"
孤独 0.3
自由 0.1
迷惘 0.1
坚强 .08
绝望 .05
...
"事物"
生命 0.3
路 0.2
夜 0.1
天空 .04
孩子 .03
...
"行动"
爱 0.2
碎 0.1
哭 0.1
死 .07
飞 .01
...
0.1
0.5
0.4
p(z|d)
d
坚强的孩子
依然前行在路上
张开翅膀飞向自由
让雨水埋葬他的迷惘
...
w
$p(w|d) = \sum_z p(w|z)p(z|d)$

LDA与整群分析的区别

- 整群分析：每个文章只归属一个主题；
- LDA：每个文章可能归属一组混合主题；
- 每个词归属于一个主题。

Topic Models的类别

- HLDA（Hierarchical Latent Dirichlet Allocation），无监督分层LDA模型；
- Labeled LDA 有监督不分层topic model；
- HLLDA（ Hierarchical Labeled Latent Dirichlet Allocation 有监督分层topic model。
- Author-Topic Model 等等

语料库（2074个经济报道）

the new york times said editorial for tuesday jan tuesday the same coins and banknotes can used buy cup coffee and the morning paper amsterdam lisbon helsinki naples dublin and dresden the franc mark lira and other currencies are disappearing vanquished the euro nothing quite like this changeover has ever taken place all goes according plan some billion worth newly minted euros will enter into circulation tuesday every bank and retail establishment from the southwestern corner portugal where the atlantic meets the mediterranean finland northernmost tundra over the next two months the euro will displace the traditional currencies used million people countries day planners call represents more than breathtaking logistical challenge and financial milestone for europe also day great political significance with euros circulation the process european integration first championed half century ago visionary french statesman named jean monnet acquires its most potent and tangible symbol the euro has been virtual currency since that when the euroland countries surrendered control their monetary policy the european central bank pegged the value their currencies the euro and made the official currency for all interbank transactions the euro has proved success even though has failed meet exuberant predictions that would instantly rival the dollar global currency its value against the dollar has actually declined percent since britain denmark and sweden three european union members that have not yet agreed adopt the euro the prevailing sense among experts that they must inevitably join the monetary union the cost doing business across the european union theory single market even before had single currency has dropped significantly the euro has made easier for european companies merge and raise capital efficient euro bond and security markets the fiscal discipline required nations before being allowed join the european monetary union forced many them undertake painful but necessary reforms the result has been low inflationary environment and leveling economic conditions though the european central bank still struggles account for regional differences when setting single interest rate for such large economy countries have until the end february retire their marks francs lire and other currencies and they are doing different times within the period the transition costs are significant retailers operating the front lines the change have been forced store five times much cash usual reconfigure millions vending machines and hire extra cashiers for the confusion likely reign the register coming days officials must make sure the changeover goes smoothly cash shortages widespread reports price gouging items are repriced could undermine the new currency unlike the common currencies offered earlier eras conquerors like charlemagne trading states like 13th century florence the euro will rely largely the public confidence for its strength and durability makes its debut there every reason think the euro here stay

Topic Model分析过程

Chose Input File: C:\testdata_news_economy_2073docs.txt

Importing and Training...this may take a few minutes depending on collection size.

Importing from: C:\testdata_news_economy_2073docs.txt.

[--remove-stopwords, true, --preserve-case, false, --input, C:\testdata_news_economy_2073docs.txt, --output, C:\topic-input.mallet, --keep-sequence]

[--num-iterations, 200, --num-top-words, 11, --doc-topics-threshold, 0.05, --input, C:\topic-input.mallet, --num-topics, 10, --output-state, C:\output_state.gz, --output-topic-keys, C:\output_topic_keys, --output-doc-topics, C:\output_doc_topics.txt]

Coded LDA: 10 topics, 4 topic bits, 1111 topic mask

max tokens: 4019

total tokens: 1292313

<10> LL/token: -9.96742

<20> LL/token: -9.48918

<30> LL/token: -9.37702

<40> LL/token: -9.32823

Topic Model分析结果（200次迭代）

1.	264	year tax economy economic billion budget percent government spending recession
2.	197	york times city jan nyt year feb team show music
3.	226	united world states country government american people japan china south
4.	219	people year million job work years state school city university
5.	361	percent company companies market year stock business economy sales quarter
6.	146	enron company energy lay officials accounting public business chairman house
7.	178	news service undated york times move stories washington receive nyt
8.	244	bush president house administration security political white war democrats party
9.	104	power palestinian people back car years small time make world
10.	134	atlanta journal constitution news cox moved beach palm service editor

TOPIC 1: year tax economy economic billion budget percent government spending recession ...

top-ranked docs in this topic (#words in doc assigned to this topic)

2.	(711)	doc 98
3.	(662)	doc 56
4.	(659)	doc 183
5.	(658)	doc 54
6.	(655)	doc 101
7.	(647)	doc 55
8.	(568)	doc 638
9.	(477)	doc 1728
10.	(438)	doc 1208
11.	(396)	doc 736
12.	(392)	doc 737
13.	(384)	doc 1004
14.	(371)	doc 1003
15.	(370)	doc 433
16.	(369)	doc 1177

DOC : doc 1

Top topics in this doc (% words in doc assigned to this topic)

---

the new york times said editorial for tuesday jan tuesday the same coins and banknotes can used buy cup coffee and the morning paper amsterdam lisbon helsinki naples dublin and dresden the franc mark lira and other currencies are disappearing vanquished the euro nothing quite like this changeover has ever taken place all goes according plan some billion worth newly minted euros will enter into circulation tuesday every bank and retail establishment from the southwestern corner portugal where the...

(30%)	year tax economy economic billion budget percent government spending recession ...
(22%)	united world states country government american people japan china south ...
(17%)	percent company companies market year stock business economy sales quarter ...
(8%)	bush president house administration security political white war democrats party ...
(6%)	news service undated york times move stories washington receive nyt ...
(5%)	york times city jan nyt year feb team show music ...

DOC : doc 56

Top topics in this doc (% words in doc assigned to this topic)

attention editors this article from the newhouse news service report thursday january available separate buy for information about purchasing the article contact debra weydert the new york times syndicate fax mail weyded nytimes com art advisory with two photos chilean vineyards nns2 and margarita aguilera nns3 for information about purchasing art call optional trim words alex pulaski newhouse news service santiago chile fertile fields flank the pan american highway cuts through the central vall...

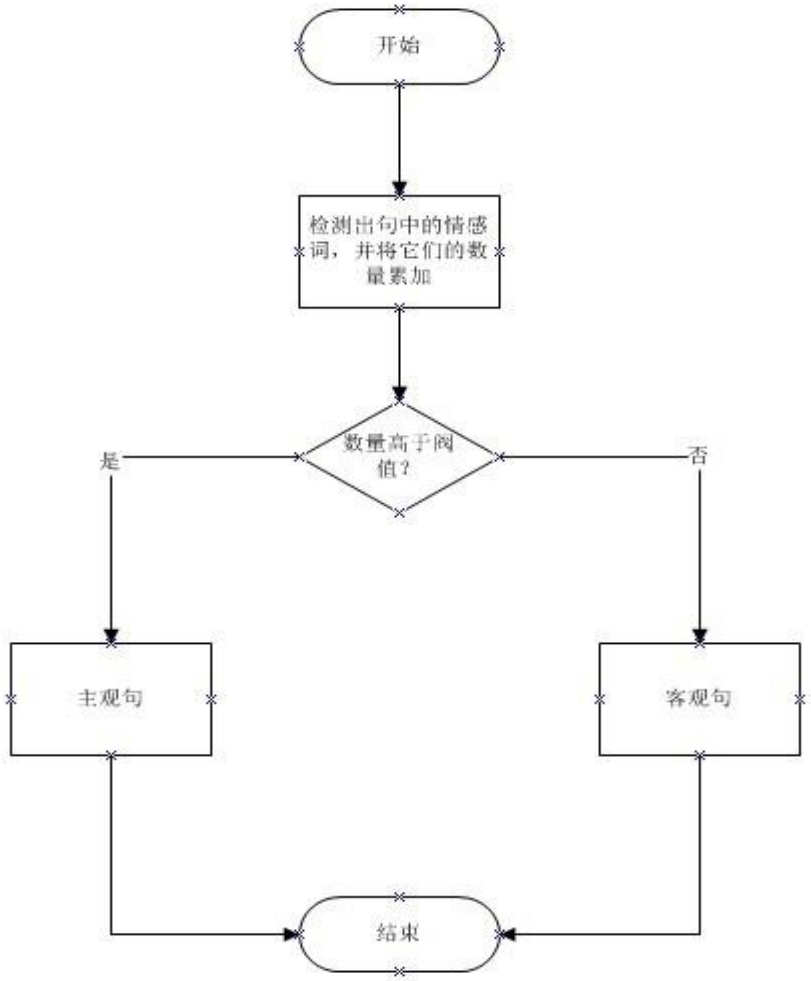
(79%)

year tax economy economic billion budget percent government spending recession ...

情感分析

- 文本情感分析（Sentiment Analysis）：通过挖掘和分析文本中的立场、观点、情绪等观点，对文本的情感倾向做出类别判断。
- 传统上，更多关注文本中的客观信息的挖掘，如将文本分类、文本检索、词频分析等；
- 文本所蕴含的主观信息在政治学、心理学等社会科学研究中也至关重要；
- 应用对象：评论、社交网络发言、政治宣言等

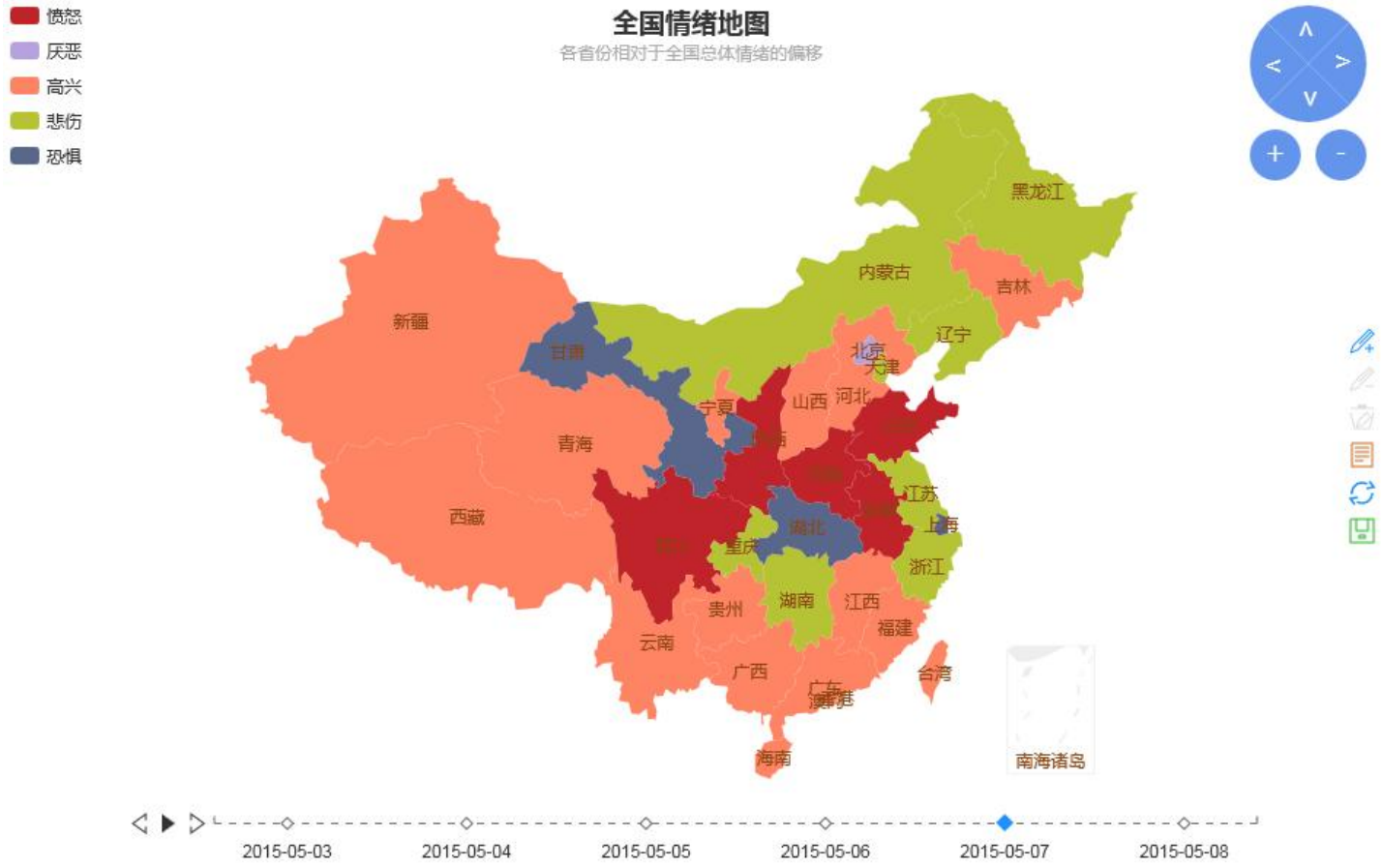
情感分析的原理



情感分析步骤

- **获取文本**
- **分词**
- **建立建模：**向量空间模型
- **选择情感词典：**中文情感词典如“hownet”
- **特征提取：**判断文本极性和提取特征词
- 设定阀值，判断文本是主观句还是客观句；
- 运用概率统计的方法，判断文本是肯定还是否定的情绪

情绪地图


愤怒
厌恶
高兴
悲伤
恐惧
全国情绪地图
各省份相对于全国总体情绪的偏移
黑龙江
内蒙古
吉林
新疆
辽宁
甘肃
北京
天津
山西
河北
宁夏
青海
江苏
西藏
上海
浙江
湖南
江西
贵州
福建
云南
广西
台湾
海南
南海诸岛
2015-05-03
2015-05-04
2015-05-05
2015-05-06
2015-05-07
2015-05-08

语义网络分析

- 语义网络分析是一种用于支持知识建构、分析推理和探索性分析的可视化文本分析方法。
- 自然语言过程及认知科学领域研究中的一个概念，70年代初由西蒙（R.F.Simon）提出；
- 可以表达复杂的概念及其之间的相互关系，是一个有向图，其顶点表示概念，而边则表示这些概念间的语义关系，从而形成一个由节点和弧组成的语义网络描述图。

- 从非结构化文本数据中自动提取语义网络；
- 语义网络利用社会网络分析方法获得对文本信息在定量和定性上的理解；

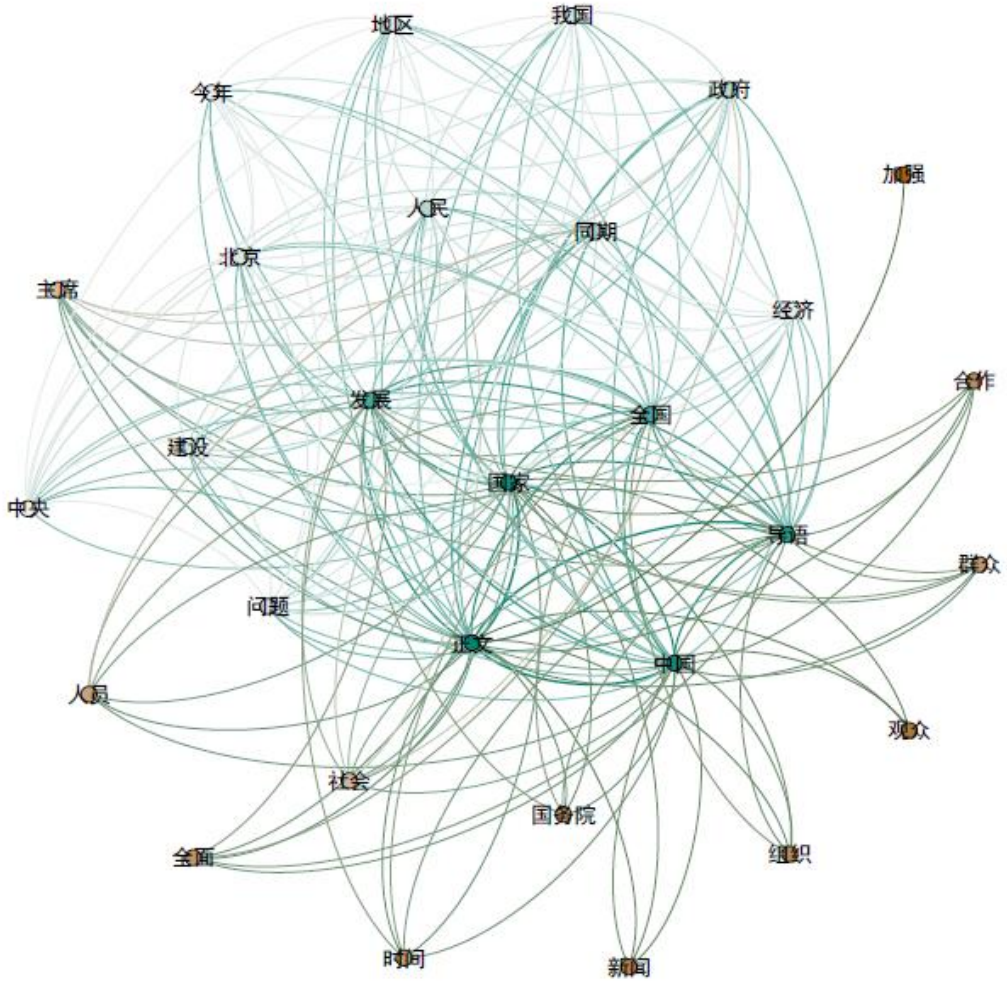
社会网络分析

- 社会网络分析是理解社会网络中结构、互动及策略位置的方法。
- 社会网络：人、组织、地方之关系的特征
- 语义网络分析：结合文本挖掘工具和网络分析来发现语言使用的社会结构特征

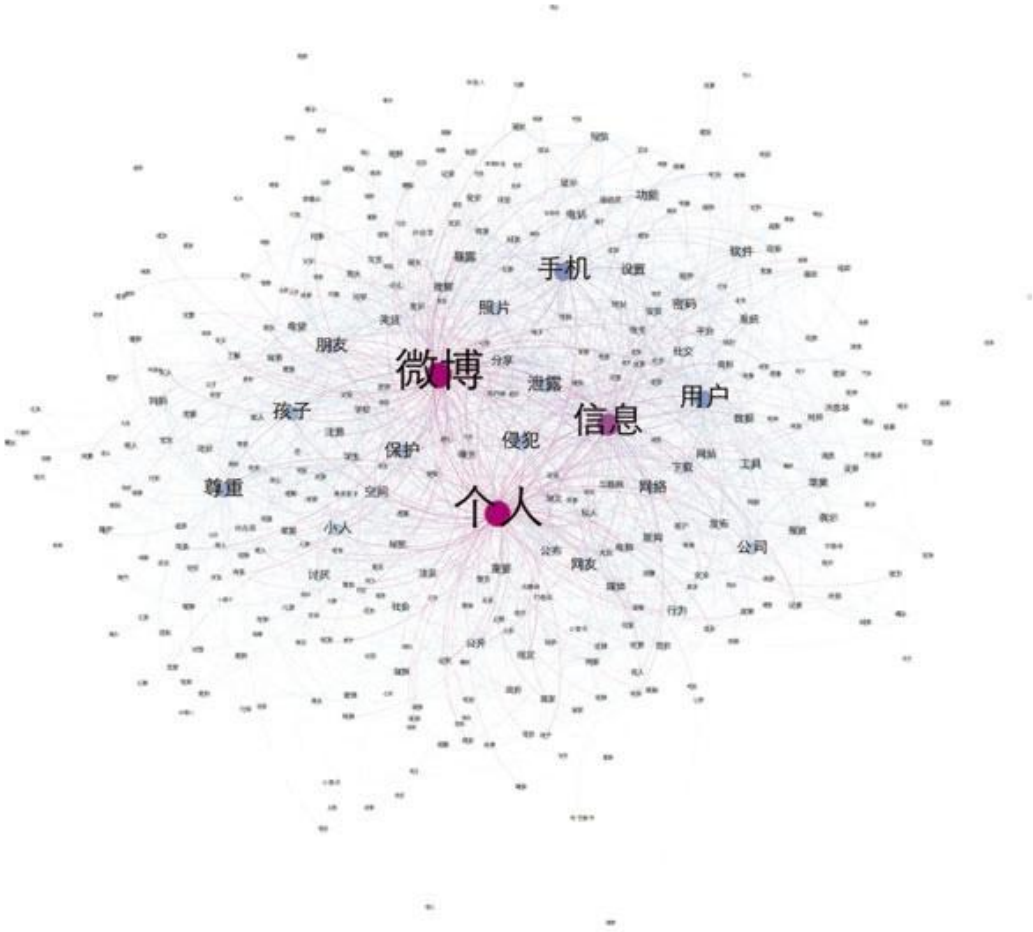
- 语义网络分析源于认知科学，认为人类记忆存在结构化意义系统（Collins & Quillian, 1972）
- 语义网络分析认为，词频、共现关系、词语间距离等元素有助于研究者探索文本中蕴含的意义体系，呈现集体认知结构（Danowski, 1993; Doerfel, 1998）

- 强化对语义结构的分析；
- 理解词语间关系；
- 识别话语体系；
- 分析概念的结构；
- 文本分类；

新闻联播的语义网络


地区
我国
今年
政府
加强
人民
同期
北京
主席
经济
合作
发展
全国
建设
国家
中央
会谈
群众
问题
北京
中国
人民
观众
社会
国务院
全面
组织
时间
新闻

新浪微博中“隐私”的语义网络



文本数据分析举例

- 选择性政府回应
- 开源文本分析工具的应用

Selective Responsiveness:
Online Public Demands and Government Responsiveness in China

Tianguang Meng
Tsinghua University

Motivation

- The continuing responsiveness of government to its citizens' preferences is considered as one of the defining characteristics of democracy.
  - Opinion-policy Nexus (Page and Shapiro,1983; Soroka and Wlezien,2010)
  - Mature Democracies (Soroka and Wlezien,2010);
  - Emerging Democracies (Roberts and Kim, 2011);

Motivation

- A growing body of research shows China, more or less, are also responsive to societal demands.
  - Representation of voters' demands in legislature (Truex 2014);
  - Willingness to incorporate public suggestions into policy making (Meng etc 2014);
  - Necessary information and help given to the needed (Distehorst and Hou 2014, Chen etc 2015).

Theoretically,

- Why local government respond?
  - The institutional foundation (Chen etc 2015);
  - Local social economic conditions (Distelhorst and Hou 2015) ;
  - Local state-society relations (Meng etc 2014).
- The second dimension? Individually
  - The nature of the issues (being submitted): Single issue or multiple issue, Economic issue or non-economic issue;
  - Social identity: household registration system, local residents and migrants, collective or individual action

Methodologically,

- Field experiment
  - Designed treatments;
  - Limited policy issues;
  - Cross-sectional design.
- Data science approach
  - Collecting the “digital footprint” of government-citizen online interactions;
  - Applying automated text analysis;
  - Longitudinal dataset that includes variations in both petitions and responsiveness by time, space, contents and individual characteristics.

Responsiveness?

- Lack of election → no incentive
- Less civic involvement → unclear knowledge and preferences
- Lack of institutional channels → limited expression and interaction
- Internet: Pressure, Participation, and Informal Channels

How Internet Changes Responsiveness?

- Online participation exerts influences on both people and their government.
- Citizens regardless of status have more equal opportunities to express opinions through internet. Magnify the pressures from below for governments to respond.
- Governments also find it a more efficient, convenient and low-costly way to respond to its citizens with new internet technologies.
- Online political forum as an institutional innovation in China.

Who Gets Responded?

- Unequal responsiveness
  - Wealth, Race or ethnicity bias.
- New sources for selective responsiveness
  - Hukou system and local bias;
  - Collective petition and individual petition
- Hypotheses
  - demands made by **local citizens** are generally more responded.
  - requests representing **collective demands** are more responded, compared to those expressing personal complaints.

Responsiveness to What?

- Ideally government should be responsive to what people desired.
- Yet many studies show responsiveness exists only in some policy domains.
- Priority issue in China
- Hypotheses
  - Policy demands related to **economic performance** are more likely to get responded.
  - Demands with **single-issue** or less demanding in cost and coordination are more likely to get responded, compared to the multiple-tasks issue.

Data

- The nationwide online political forum—“Local Leadership Message Board” (LLMB)
  - Supported by the Central official media people.cn;
  - Beyond the control of local governments;
  - Transparent and open access;

Data

- Data source
- On trial for 2006 and 2007 and formally running since 2008;
- Different local level governments: provincial, prefecture city, and county
- Scrapying over 200,000 records on provincial top leaders from 2008 to early 2014

Data Generating

- Submitting a demand by citizen;
- Realname or anonymous;
- Assign tasks;
- Local government respond;

【其他 求助】离婚后户口无法回迁 | 已回复

网友：匿名网友 2015-11-28 21:47

韩书记，你好！我与前妻三年前离了婚，而我的户口在她家，至今无法回迁到娘家，因娘家农村宅基地已动迁，安置的房子没产权证，我户籍地是松江区，现在户口在金山区前妻家里，两边的派出所都说不能办，因为前妻不同意我再婚后生的孩子户口随我放在前妻家，难道就没有一点办法了吗，我现在的妻子是外地的，难道把我小孩的户口报到外地去吗？这政策是要逼死人吗？

官方回复

回复单位：上海市公安局 2016-03-03 10:18

尊敬的网友：

您好！经查，您和外省女性再婚后生育的子女落户问题，根据《上海市常住户口管理规定》[沪公发（2010）370号]第八条规定，婴儿可以在本市父或母户口所在地公安派出所办理出生登记，并统一登记为非农业户口。您要求将新生子女户口直接申报到其他直系亲属处，并无政策依据。

您父母户籍地为上海市松江区泖港镇曙光村1018号，此地址房屋已拆迁，您父母居住地为上海市松江区泖港镇五库北库路116弄19号402室，系集资房无产权证，两处地址都无法让您回迁户口，您的户口性质为农业户口，所以无法迁入辖区公共户内，而您本人亦不愿意将户口从金山区前妻处迁出转为非农业户口。其又不想将小孩的户口随母报到外地。

故此，按照现行的户籍政策，您子女的户口无法直接在泖港所申报。

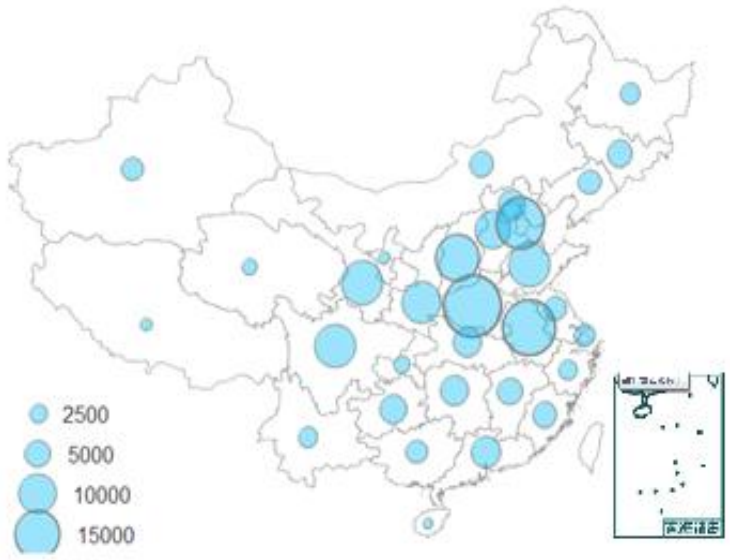
2016年1月19日，分局泖港派出所民警电话答复您，按照现行的户口政策，您子女的户口无法直接在泖港派出所申报，待户口政策允许时，会及时通知您。

Data

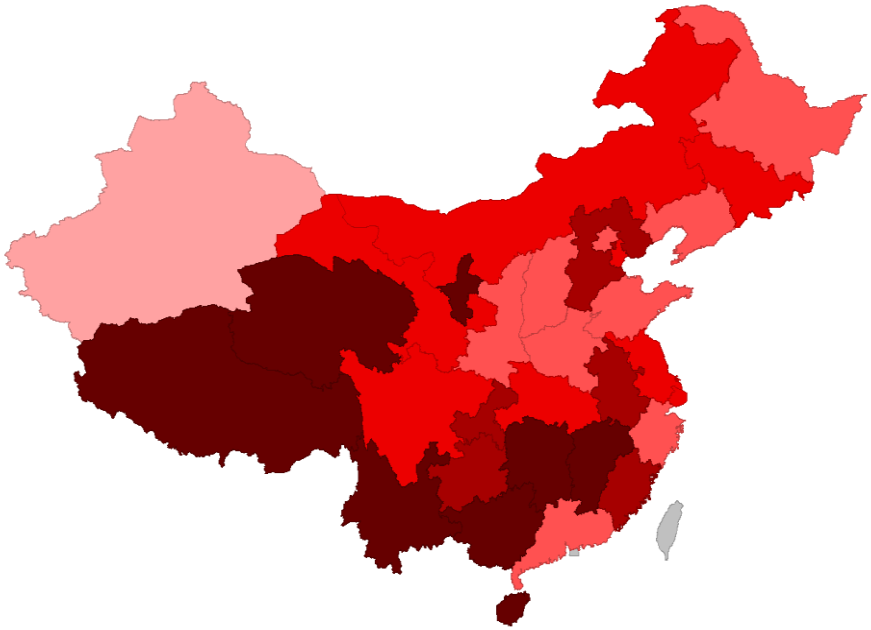
Year	Demands to CCP Secretaries		Demands to Governors		Sum
	Cases	Percentage	Cases	Percentage	
2008	11,222	71.7	4,438	28.3	15,660
2009	18,584	71.6	7,355	28.4	25,939
2010	21,131	68.5	9,697	31.5	31,175
2011	18,754	63.3	10,892	36.7	30,018
2012	21,311	64.7	11,607	35.3	32,919
2013	27,317	63.2	15,875	36.8	43,192
Early 2014	22,772	68.3	10,547	31.7	33,321
Total	141,091	66.7	70,441	33.3	211,502

Data

Distribution of Online Public Demands



Nonlocal Petitions



Examples of Public Demands

- Example 1: Collective demands related to employment from rural teachers

Dear Secretary Sun,

I am an ordinary rural teacher in Tianjin. As the price level keeps rising and the housing price continuously grows, **our wages are relatively low**. We earnestly request you to further pay attention to the issue of increasing teachers' salaries. Thank you!

Examples of Public Demands

- Example 2: Individual demand related to agriculture from a local citizen

Dear Secretary Zhang Qingli,

I am a peasant at Xinmatou Town, Qiu County, Handan City. Eight mu of **arable land**, which belongs to our family, has been compulsorily collected by the township government in an illegal way of Yizu Daizheng (paying "rents" for land expropriation to avoid regulations). Please look into this matter, our leaders!

Examples of Public Demands

- Example 3: Individual demand related to business from a nonlocal citizen

Dear Governor Wang Guosheng,

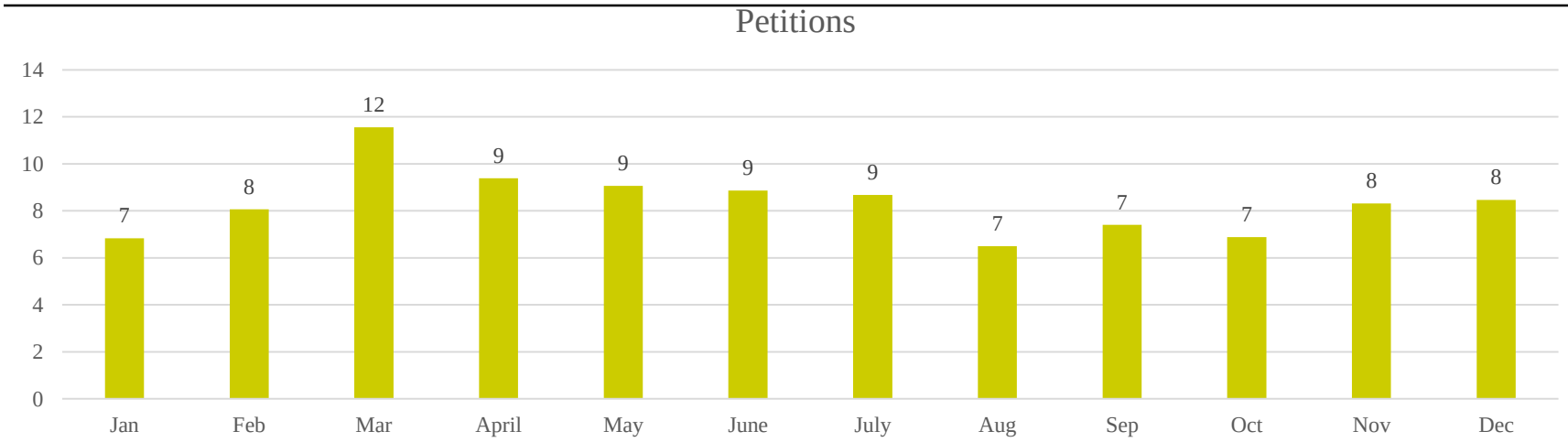
I am a businessman from Jiangsu Province making investments in Yichang City (Hubei Province). I have been unfairly treated in your area and hold a different opinion toward the practices of related government branches. I want to report to you this kind of administrative omission and please give me a chance.

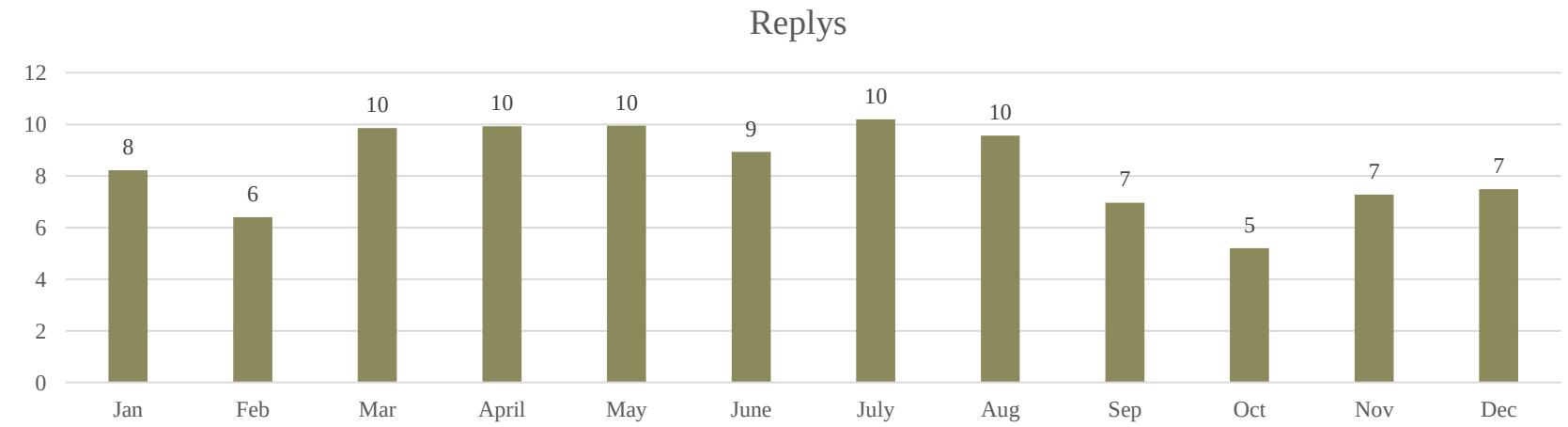
Examples of Public Demands

- Example 4: Individual demand related to removal and resettlement from a local citizen

I am a villager from Linjiang Village, Wangyang Town, where the demolition work for Huateng Coking Plant project is being implemented. The measure of indemnity follows the 2008 standard and the local officials say there are not any new policies since 2008. Is this true? Is it reasonable that our buildings with brick and concrete structure are compensated at the rate of 500 yuan per square meter, while we should pay 760 to 840 yuan (per square meter) to buy our resettlement housing back?

Data


Petitions
14
12
10
8
6
4
2
0
7
8
12
9
9
9
9
7
7
7
8
8
Jan
Feb
Mar
April
May
June
July
Aug
Sep
Oct
Nov
Dec


Replys
12
10
8
6
4
2
0
8
6
10
10
10
9
10
10
7
5
7
7
Jan
Feb
Mar
April
May
June
July
Aug
Sep
Oct
Nov
Dec

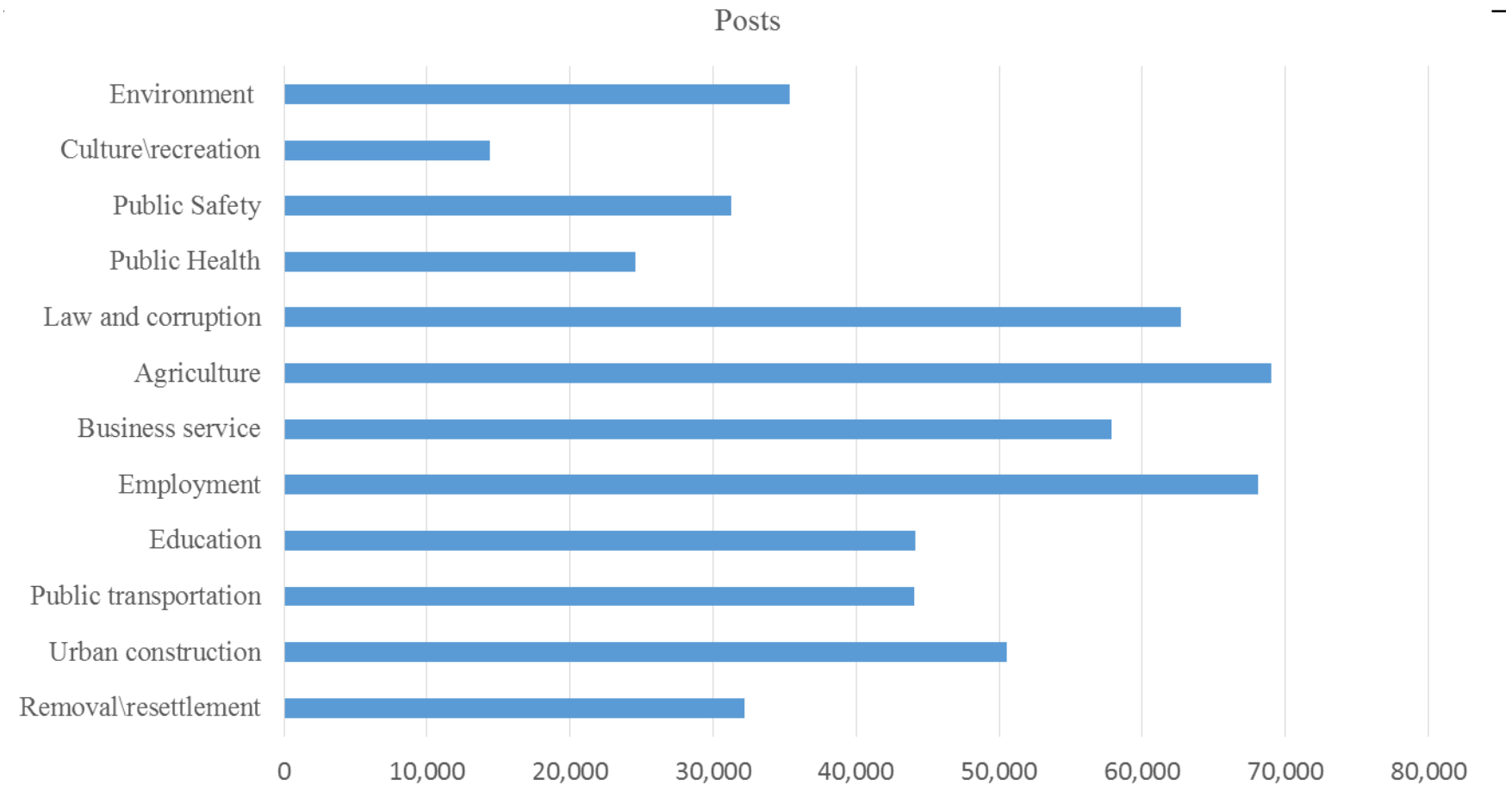
Methods

- Automated text analysis
  - Supervised approach
  - RAs read and code the records in the randomly selected training set
  - Automatically classify the larger records with computer
  - Randomly testing set of 2,000 posts, and classified manually for examining the consistency.
- Sentiment analysis

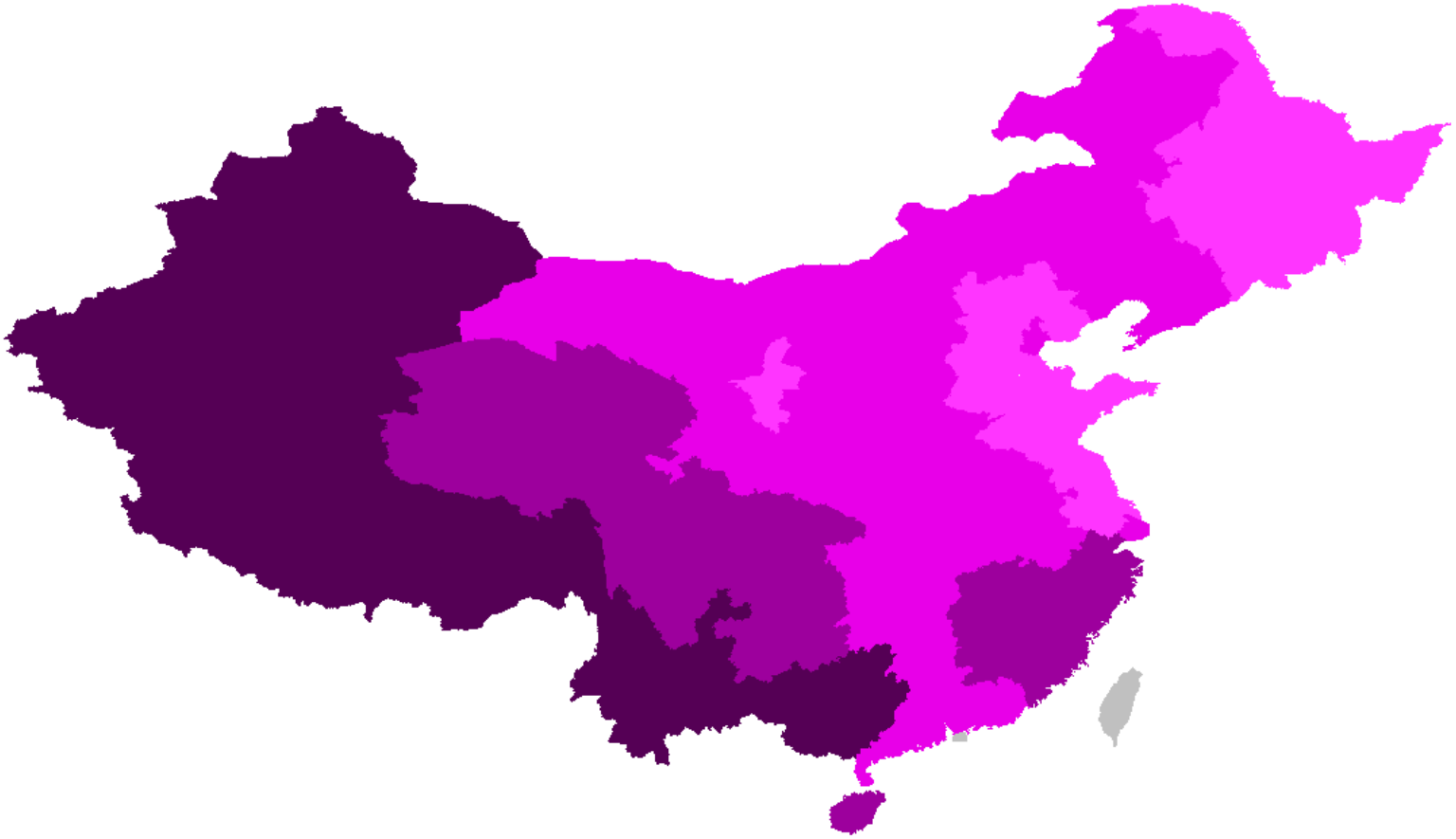
Methods

- Logistic regression
  - DV：Reply or not
  - IV：Social identity, policy issue,
  - CV：length of petition, sentiment, real-name petition, party secretary or governor, local political and economic conditions, year and province FE
- Duration regression (Cox proportional hazards model)
  - DV：Days for reply
  - IV：Social identity, policy issue,
  - CV：length of petition, sentiment, real-name petition, party secretary or governor, local political and economic conditions, year and province FE

Policy Domains of Online Posts



Sentiment of Online Demands



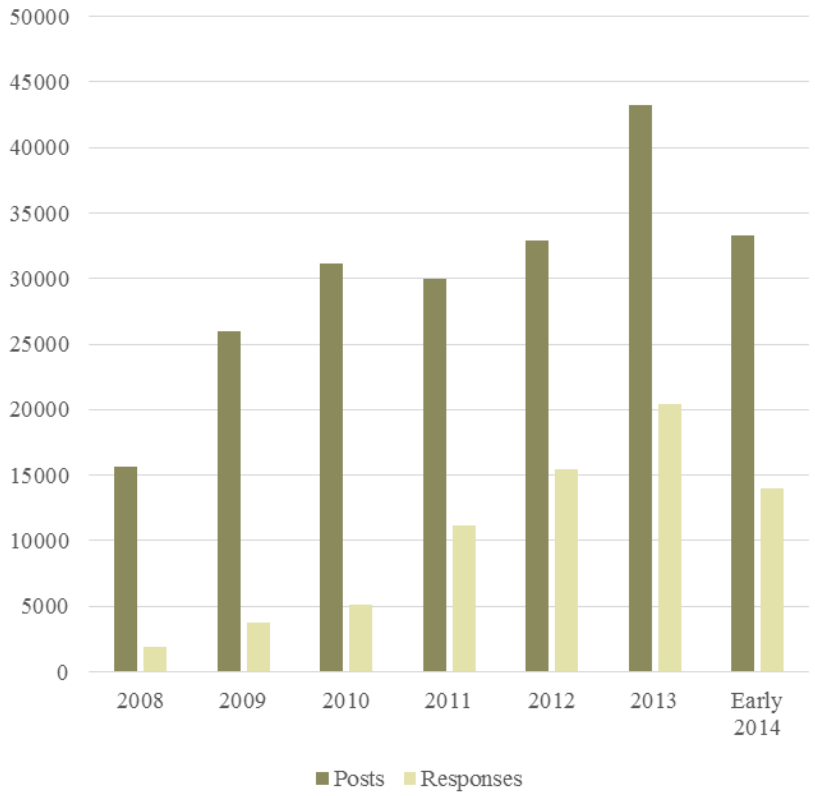
Petitions among localities

Beijing

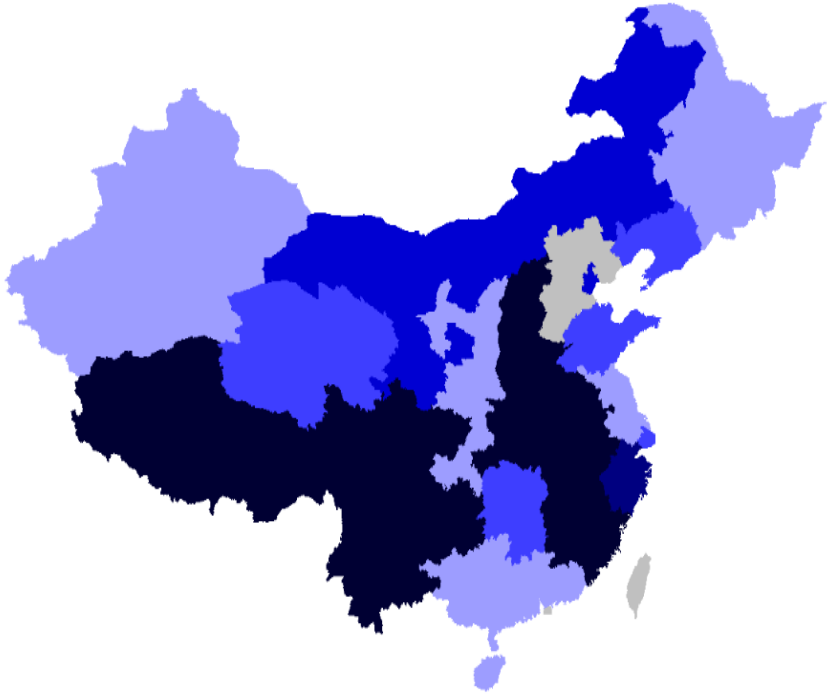
Shanghai

Level of Responsiveness

Yearly

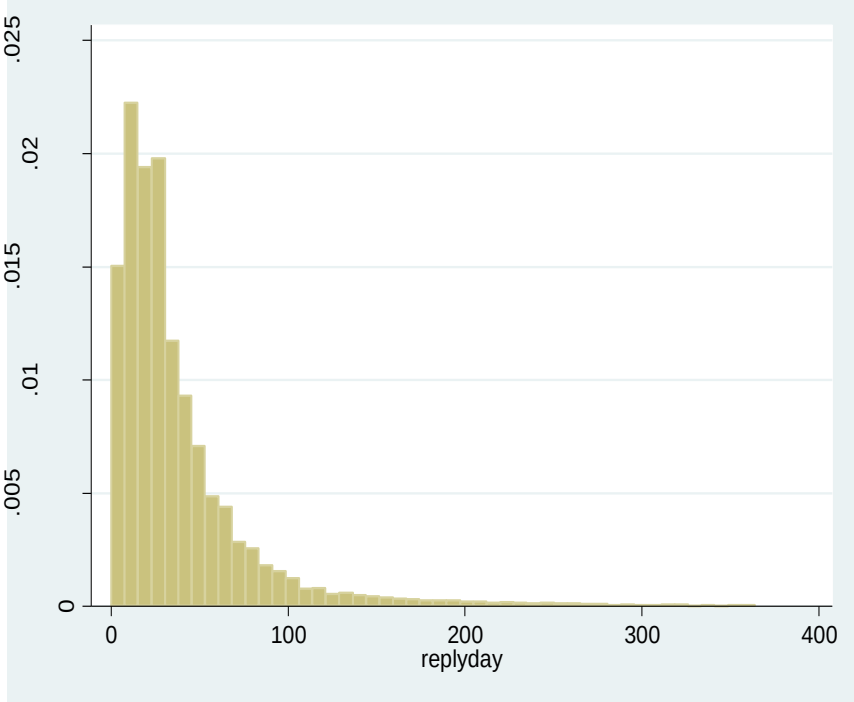


Geographically

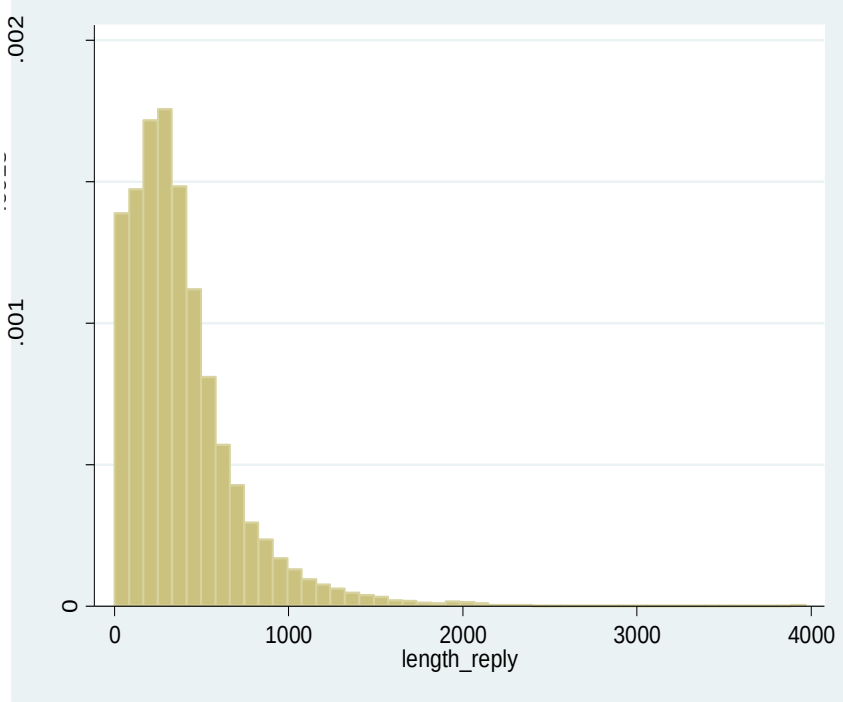


Level of Responsiveness

Days for Reply



Words for Reply



Logistic models of response to demands

	(1)	(2a)	(2b)	(3)	(4a)	(4b)	(5a)	(5b)
	Odds Ratio (Z-value)	Odds Ratio (Z-value)	Odds Ratio (Z-value)	Odds Ratio (Z-value)	Odds Ratio (Z-value)	Odds Ratio (Z-value)	Odds Ratio (Z-value)	Odds Ratio (Z-value)
Collective demand		**1.19***(12.83)**	**1.09***(3.78)**		**1.49***(24.85)**	**1.19***(6.61)**	**1.49***(24.5)**	**1.19***(6.32)**
Individual demand		**0.83***(-10)**	**0.83***(-6.43)**		**0.82***(-9.34)**	**0.84*** (-5.39)**	**0.82***(-9.27)**	**0.84*** (-5.43)**
Nonlocal IP		**1.05** (3.03)**	**1.06* (2.46)**		**0.97 (-1.84)**	**1.01 (0.27)**	**0.97*(-2.07)**	**1.00 (0.16)**
Nonlocal citizen			**0.95 (-1.52)**			**0.90** (-2.68)**		**0.91** (-2.40)**
Local citizen			**1.11*** (3.77)**			**1.64*** (12.57)**		**1.64*** (12.50)**
Policy Domains								
Removal\resettlement				**0.89*** (-6.00)**	**0.87*** (-6.66)**	**0.86*** (-5.11)**	**0.88*** (-6.44)**	**0.86*** (-5.05)**
Urban construction				**1.95*** (36.14)**	**1.77*** (30.19)**	**1.70*** (17.6)**	**1.75*** (29.46)**	**1.70*** (17.3)**
Public transportation				**1.24*** (12.22)**	**1.26*** (12.94)**	**1.27*** (8.02)**	**1.26*** (1269)**	**1.27*** (8.04)**
Education				**1.01 (0.57)**	**0.89*** (-6.07)**	**0.97 (-1.19)**	**0.88*** (-6.53)**	**0.96 (-1.50)**
Employment				**0.72*** (-16.55)**	**0.72*** (-15.98)**	**0.77*** (-8.27)**	**0.72*** (-15.76)**	**0.77*** (-8.14)**
Business service				**1.05* (2.15)**	**0.96 (-1.76)**	**0.96 (-1.23)**	**0.96 (-1.89)**	**0.96 (-1.24)**
Agriculture				**1.03 (1.74)**	**0.97 (-1.92)**	**1.04 (1.68)**	**0.97 (-1.69)**	**1.05 (1.90)**
Law and corruption				**1.06*** (3.49)**	**1.06** (3.46)**	**1.08** (3.08)**	**1.06** (3.34)**	**1.08** (2.96)**
Public Health				**1.32*** (12.1)**	**1.37*** (13.64)**	**1.31*** (7.77)**	**1.36*** (13.42)**	**1.30*** (7.63)**
Public Safety				**1.17*** (7.46)**	**1.22*** (9.44)**	**1.12** (3.42)**	**1.22*** (9.51)**	**1.13** (3.52)**
Culture\recreation				**1.02 (0.78)**	**1.05 (1.73)**	**1.10 (1.91)**	**1.05 (1.82)**	**1.10*(1.98)**
Environment				**0.00*** (-21.15)**	**0.00*** (-21.27)**	**0.00*** (-9.69)**	**0.00*** (-21.05)**	**0.00*** (-9.70)**
Ln(Length of petition)	1.27*** (29.22)	1.27*** (26.68)	1.07*** (4.04)	1.47*** (34.98)	1.44*** (32.04)	1.23*** (9.85)	1.45*** (32.40)	1.23*** (9.69)
Sentiment	0.997***(-11.65)	0.997*** (-12.33)	0.998*** (-4.66)	0.998*** (-7.00)	0.998*** (-6.1)	0.999* (-2.25)	0.998*** (-6.16)	0.999* (-2.24)
Real-name petition	0.95** (-2.78)	0.96* (-2.2)	0.98 (-0.72)	0.92*** (-3.83)	0.93** (-3.15)	0.97 (-0.93)	0.93** (-3.14)	0.96 (-1.23)
Party secretary	1.12*** (8.29)	1.12*** (7.99)	1.14*** (6.15)	1.06*** (3.65)	1.06*** (3.65)	1.11*** (4.35)	1.06*** (3.66)	1.10*** (4.02)
Urban Rate							1.53*** (35.41)	1.64*** (24.7)
Netizen Percent							0.00*** (-28.2)	0.00*** (-19.0)
Tenure of Governor (year)							1.06** (10.07)	1.09*** (9.45)
Tenure of Secretary (year)							1.07*** (13.39)	1.08*** (9.91)
Constant	0.00*** (-12.52)	0.00*** (-12.57)	0.00*** (-9.94)	0.00*** (-12.92)	0.00*** (-12.85)	0.00*** (-10.34)	0.00*** (-28.58)	0.00*** (-22.17)
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Province	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Log likelihood	-78366.6	-77168.6	-30156.63	-63963.2	-63581.8	-24828.8	-62668.2	-24369.7
Pseudo R2	0.4142	0.4163	0.3877	0.5139	0.5167	0.4949	0.5237	0.5043
N	211,117	208,273	73,781	207,554	207,493	73,663	207493	73,663

Duration models of response to demands

	(1)	(2a)	(2b)	(3)	(4a)	(4b)	(5a)	(5b)
	Hazard Ratio (Z-value)	Hazard Ratio (Z-value)	Hazard Ratio (Z-value)	Hazard Ratio (Z-value)	Hazard Ratio (Z-value)	Hazard Ratio (Z-value)	Hazard Ratio (Z-value)	Hazard Ratio (Z-value)
Collective demand		**1.11*** (13.09)**	**1.06*** (3.99)**		**1.22*** (22.53)**	**1.10*** (6.18)**	**1.22*** (22.20)**	**1.09*** (6.02)**
Individual demand		**0.89*** (-9.74)**	**0.91*** (-5.48)**		**0.91*** (-8.02)**	**0.92*** (-4.48)**	**0.91*** (-8.13)**	**0.92*** (-4.58)**
Nonlocal IP		**1.02* (2.05)**	**1.02 (1.62)**		**0.98* (-1.98)**	**1.00 (0.19)**	**0.98** (-2.59)**	**1.00 (0.03)**
Nonlocal citizen			**0.96* (-2.24)**			**0.94** (-2.87)**		**0.94** (-2.85)**
Local citizen			**1.11*** (6.3)**			**1.25*** (11.94)**		**1.24*** (11.76)**
Policy Domains								
Removal\resettlement				**0.94*** (-5.57)**	**0.94*** (-5.94)**	**0.95** (-3.27)**	**0.94*** (-5.74)**	**0.95** (-3.22)**
Urban construction				**1.35*** (32.22)**	**1.29*** (26.2)**	**1.24*** (14.46)**	**1.28*** (25.93)**	**1.25*** (14.51)**
Public transportation				**1.09*** (8.81)**	**1.09*** (9.71)**	**1.10*** (6.13)**	**1.09*** (9.47)**	**1.10*** (6.15)**
Education				**0.99 (-0.67)**	**0.94*** (-6.43)**	**0.98 (-1.4)**	**0.93*** (-6.68)**	**0.98 (-1.65)**
Employment				**0.84*** (-16.26)**	**0.84*** (-15.84)**	**0.88*** (-7.83)**	**0.84*** (-15.71)**	**0.88*** (-7.67)**
Business service				**0.99 (-0.58)**	**0.96*** (-3.62)**	**0.96* (-2.24)**	**0.96*** (-3.59)**	**0.96* (-2.22)**
Agriculture				**1.01 (0.84)**	**0.98* (-2.09)**	**1.03* (2.36)**	**0.98* (-1.98)**	**1.03* (2.35)**
Law and corruption				**1.03** (2.73)**	**1.02* (2.56)**	**1.03* (1.98)**	**1.02* (2.28)**	**1.03 (1.95)**
Public Health				**1.11*** (8.68)**	**1.13*** (9.88)**	**1.11*** (5.55)**	**1.13*** (9.78)**	**1.10*** (5.51)**
Public Safety				**1.05*** (4.75)**	**1.07*** (6.18)**	**1.04* (2.37)**	**1.07*** (6.25)**	**1.05** (2.74)**
Culture\recreation				**0.998 (-0.15)**	**1.01 (0.7)**	**1.05 (1.8)**	**1.01 (0.35)**	**1.04 (1.44)**
Environment				-	-	-	-	-
Ln(Length of demand)	1.14*** (26.58)	1.13*** (23.79)	1.01 (1.34)	1.21*** (31.51)	1.19*** (28.58)	1.08*** (6.82)	1.20*** (28.94)	1.08*** (6.88)
Sentiment	0.998*** (-12.1)	0.998*** (-12.52)	0.999*** (-4.26)	0.999*** (-9.57)	0.999*** (-8.48)	0.999** (-3.47)	0.999*** (-8.48)	0.999** (-3.36)
Real-name demand	0.97** (-2.67)	0.98* (-2.21)	0.98 (-1.09)	0.95*** (-4.33)	0.96*** (-3.66)	0.98 (-1.35)	0.96*** (-3.72)	0.97 (-1.64)
Party secretary	1.25*** (25.67)	1.25*** (25.48)	1.24*** (15.96)	1.19*** (20.1)	1.19*** (19.99)	1.19*** (12.93)	1.19*** (20.25)	1.19*** (12.98)
Urban Rate							1.30*** (31.68)	1.39*** (24.32)
Netizen Percent							0.00*** (-23.1)	0.00*** (-18.8)
Tenure of Governor (year)							0.97*** (-7.17)	0.99 (-1.00)
Tenure of Secretory (year)							0.95*** (-15.49)	0.96*** (-7.71)
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Province	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Log likelihood	-773283.6	-765328.2	-290433.3	-747743.6	-747329.1	-284547.2	-746413.7	-284059.3
LR Chi^2	123049.4***	122152.2***	43595.9***	146254.2***	146818.8***	53570.6***	146818.8***	53570.6***
n of subjects	209,156	206,315	73,128	205,610	205,549	73,011	205,549	73,011
n of failures	69,503	68,888	28,583	68,443	68,434	28,504	68,434	28,504

Conclusion and discussion

- Adoption and implementation of daily political interaction between citizens and government online.
- Selective nature of online responsiveness in China
  - local bias;
  - expressing demands collectively;
  - single task issue;
  - closely related to economic growth.

Conclusion and discussion

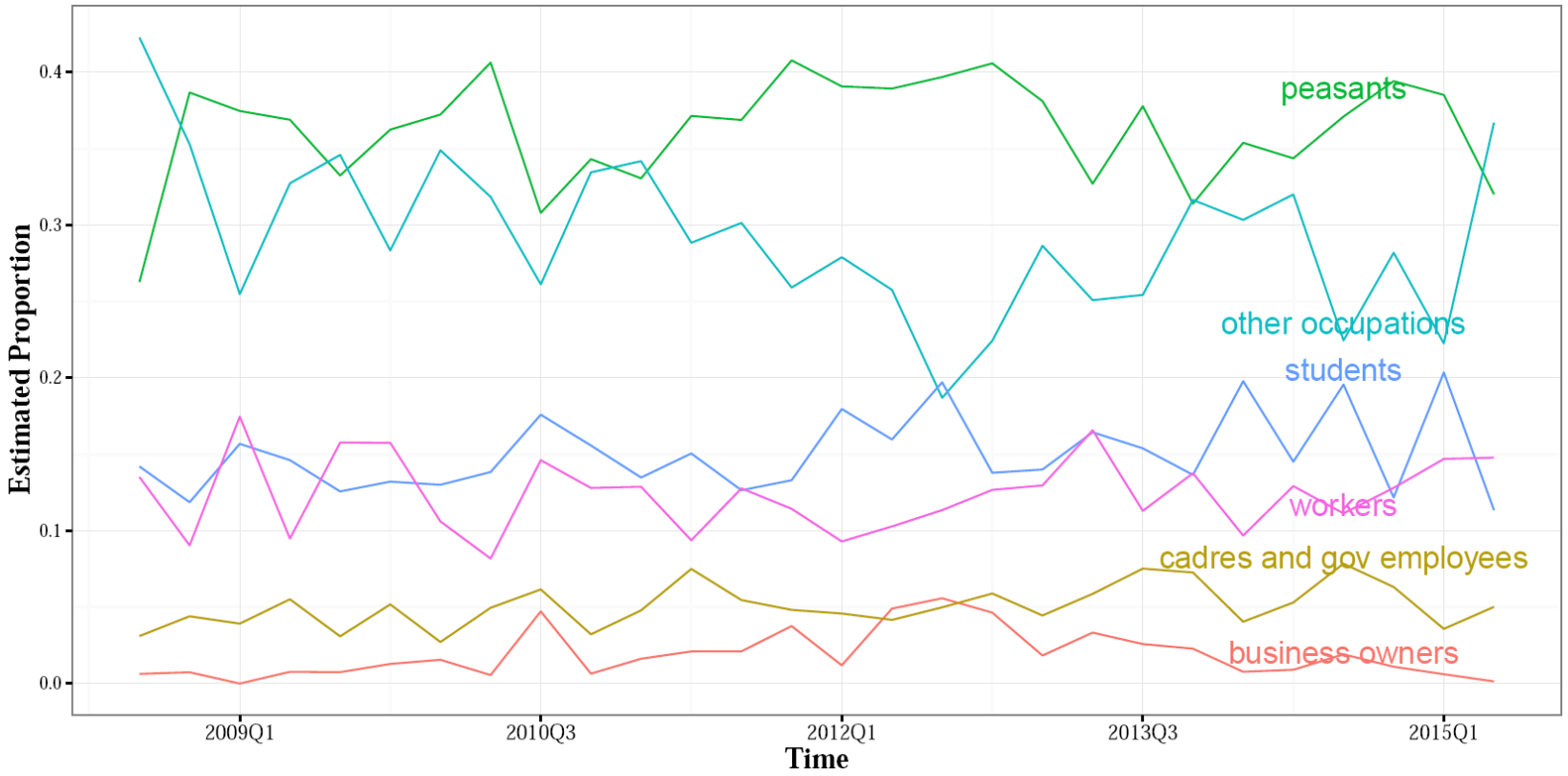
- Potentials for big data analytics in China Politics
  - More convenient, less costly and faster data collection
  - Better tools for analyzing texts, networks, and unstructured
  - Massive data allows scholars to study small area
  - Combine data science with traditional methods;
  - But big data is not panacea. The validity and reliability of measurement, selective bias, lack of standards.

Conclusion and discussion

- Future study
- More data sources
  - City and county level interaction;
  - Local platforms;
- Better measure of the quality of responsiveness
  - Exact Match between petition and reply;
  - Richness of information;
  - Real-world consequence of online interaction.

Who Participate?

- Supervised Learning with readme (Hopkins and King, 2010)



开源文本分析工具的应用

- 百度指数
- Google trend
- http://www.google.com/trends
- Google books
- https://books.google.com/ngrams

谢谢！