

北京邮电大学
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

大数据时代的管理

Management in Big Data Era




马宝君 博士 讲师

经济管理学院
电子商务中心
2014年12月29日

1

上次课程小结：推荐系统基础知识

- 推荐系统出现的背景
- 推荐系统简介
- 推荐系统的模块
 - 用户建模模块
 - 推荐对象建模模块
 - 推荐算法模块
- 推荐系统的评测指标



2

回顾：深度业务分析——组合方法及应用

- 信息检索及信息搜索服务（文本内容、链接）
- 推荐系统及产品推荐
- 情感分析及舆情监测
- 社交网络分析及关系营销
- 用户生成内容（口碑/评论/社交）分析
-




以推荐
挖掘并满足用户的潜在需求

amazon.com
PANDORA
NETFLIX
last.fm
StumbleUpon
Discover your web

3

情感分析基础

Sentiment Analysis Foundations





4

情感分析基础知识

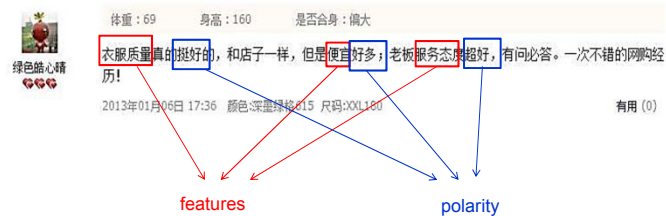
- 情感分析与观点挖掘问题定义
- 情感分析与观点挖掘的任务
 - 文档层次情感分类
 - 句子层次情感分析
 - 基于特征/方面的情感分类
- 总结

5



6

Online product review: features (属性), polarity (极性)



7

引子：事实和观点

- 互联网上两类文本信息
 - 事实
 - 观点
- 目前的搜索引擎可以查找事实 (假设它们是正确的)
 - 事实可以使用关键词、主题词表示
- 但搜索引擎不查找观点 (Opinion)
 - 观点难以使用数个关键词表示
 - 例如：人们怎样看apple手机?
- 目前的搜索排序策略不适合观点检索或搜索

8

用户生成内容：观点挖掘

- **互联网上的口水话（口碑，Word-Of-Mouth，WOM）**
 - 网民可以在电子商务平台、评论站点、论坛、博客、微博、社交网络等地方就任何事件阐述个人的经历或观点（用户生成内容，UGC）
 - 包含有价值的信息
- **我们感兴趣的是：在用户生成内容中挖掘观点**
 - 一个智能的、很有挑战性的问题
 - 随着社会化媒体的日益流行越来越重要
 - 实践中很有用



9

应用

- **商务和组织：市场情报**
 - 工商企业花费大量的金钱搜集顾客的意见和观点。
 - 顾问、调查组，等等
- **个人：在下列情况下会对他人的观点感兴趣**
 - 购买产品或使用服务
 - 寻找政治话题的观点
- **广告放置：在用户生成内容中放置广告**
 - 当用户称赞某个产品时放置广告。
 - 当用户批判某个产品时放置竞争品牌的广告
- **观点检索/搜索：提供观点的全面搜索**
-



10

情感分析与观点挖掘

- **Sentiment Analysis (SA) or Opinion Mining**
 - 针对实体、个人、问题、事件、主题和他们的属性，分析人们的观点、评价、态度和情感
 - computational study of opinion, sentiment, appraisal, evaluation, and emotion
- **重要性**
 - **观点（Opinions）**是影响我们行为的关键因素
 - 我们的观念和对现实世界的感知看法，在很大程度上依赖于其他人如何看待这个世界
 - 当需要作决策时，我们经常要寻求他人的意见
 - 个体如此、团体组织也如此
- **查找和监测网络中的观点网站和提取其中所含的信息仍然是一个艰巨的任务**

11

术语：情感、观点

- **情感（Sentiment）**
 - An attitude, thought, or judgment prompted by feeling
 - 情感更像一种感觉（feeling）
 - “I am concerned about the current state of the economy.”
- **观点（Opinion）**
 - A view, judgment, or appraisal formed in the mind about a particular matter
 - 一个人对一件事物的具体印象
 - a concrete view of a person about something.
 - “I think the economy is not doing well.”

12

观点挖掘简介

● 观点的基本要素

- **观点持有者**：对于特定对象持有特定观点的个人或组织
- **实体对象**：观点表达的作用者或对象
- **观点**：观点持有者对一个实体对象的一种看法、态度或评价

● 观点挖掘的目标: 很多 ...

- **文档层次**的目标：评论的情感分类
- **句子层次**的目标：主观或客观句子的识别，主观句子的情感分类
- **方面层次**的目标：识别实体对象特征方面，找实体对象特征方面的同义词，

13

两类观点

● 直接观点

- 关于某个对象诸如产品、事件、主题和个人的情感表达
- 例如：“the picture quality of this camera is great”
- 主观的

● 比较观点

- 表示多于一个对象的不同点或相同点的关系，通常表示一种次序
- 例如：“car x is cheaper than car y.”
- 客观的或主观的

14

观点的形式化定义

● 一个观点可以表达为一个五元组 (quintuple)

(*entity*, *aspect*, *sentiment*, *holder*, *time*)

- **entity**: target entity (or object).
- **Aspect**: aspect (or feature) of the entity.
- **Sentiment**: +, -, or neu, a rating, or an emotion.
- **holder**: opinion holder.
- **time**: time when the opinion was expressed.

● 观点挖掘 (基于特征方面) 的目标是给定一个含有观点的文档

- 找出所有的五元组 (即挖掘五元组中五个部分的对应信息)
- 或者, 解决一些更简单的问题

15

例子：观点五元组

- **Id: Abc123 on 5-1-2008** “I bought an *iPhone* a few days ago. It is such a nice *phone*. The *touch screen* is really cool. The *voice quality* is clear too. It is much better than my old *Blackberry*, which was a terrible *phone* and so *difficult to type* with its *tiny keys*. However, *my mother* was mad with me as I did not tell her before I bought the *phone*. She also thought the phone was too *expensive*, ... ”

● 可以得到例如下列的一些观点五元组

- (iPhone, **GENERAL**, +, Abc123, 5-1-2008)
- (iPhone, touch_screen, +, Abc123, 5-1-2008)
- (iPhone, voice_quality, +, Abc123, 5-1-2008)
-



16

观点挖掘的另一个策略：观点汇总

- 当存在大量的观点时，**观点汇总 (Opinion Summary)** 就很有必要了

- 传统的文本汇总方法很难直接适用 (表达观点多为较短的文本)
- 需要量化表示 (60%的正面意见 VS. 90%的正面意见)
- 主要形式 (Aspect-based opinion summary)

"I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, and wanted me to return it to the shop. ..."

基于特征的汇总:

特征1: Touch screen

Positive: 212

- The touch screen was really cool.
- The touch screen was so easy to use and can do amazing things.

...

Negative: 6

- The screen is easily scratched.

- I have a lot of difficulty in removing finger marks from the touch screen.

...

特征2: battery life

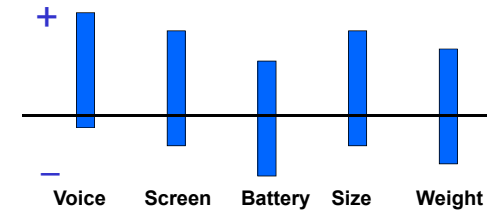
...

17

可视化汇总/比较

观点汇总

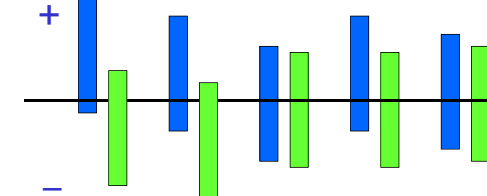
Cell Phone 1



观点比较

Cell Phone 1

Cell Phone 2



18

观点挖掘是一个难题！

- (entity, aspect, sentiment, holder, time)
 - entity – 目标对象实体：命名实体抽取 (或更多问题)
 - Aspect – 实体的特征方面：信息抽取
 - Sentiment – 情感值：情感判定
 - holder – 观点持有者：信息/数据抽取
 - time – 时间：数据抽取



- 共指消解、隐藏评论实体

- 例如：“马尔可夫模型” = “马氏模型” = “它”

- 同义词/短语匹配

- 例如：Voice = Sound quality

- 上述问题没有一个是目前完全解决好的！

19

观点挖掘的任务

- 在文档层次 (document, 例如一段评论)

- 任务：对整个评论做情感分类
- 类：正面、负面、中立
- 假设：每个文档 (或评论) 仅针对单一对象实体并且仅包含单一观点持有者的观点

- 在句子层次 (sentence, 例如一句话)

- 任务1：识别主观的/客观的句子
 - 类：客观点、主观观点
- 任务2：句子的情感分类
 - 类：正面、负面、中立
 - 假设：一个句子仅含有一个观点 (在很多情况下不成立, 我们可以进一步考虑分句或短语)



20

观点挖掘的任务（续）

● 在特征/方面层次（feature/aspect）

- 任务1: 识别和抽取被观点持有者(即评论人)评价的对象特征
- 任务2: 判定针对该特征的观点是正面的, 负面的还是中立的
- 任务3: 对特征的同义词作分组, 产生多个评论的基于特征的观点汇总



21

文档层次的情感分类

Document Sentiment Classification

22

文档层次的情感分类

● 基于观点持有者表达的**总体情感**对文档（如评论）归类

- 三分类的监督学习任务：正面(4-5星)、负面(1-2星)或中立(3星)
- 因为在观点的五元组模型中，对象O有一个特征/方面就是GENERAL，所以**情感分类**本质上是判定每个文档该观点的内容

● 与基于主题的文本分类相似但不同

- 在基于主题的文本分类中，把文本分类到各个预定义的主题上，例如政治、科学、体育等；**主题相关词**是很重要的，
- 在情感分类中，表达正面或负面观点的**情感词**或**观点词**更重要，比如great, excellent, amazing, horrible, bad, worst, 等

● 任何现有的监督学习方法都可以用于情感分类中，例如朴素贝叶斯分类、SVM等

23

基于无监督学习的文档情感分类

● 数据：从epinions.com中获取的汽车, 银行, 电影, 和旅游目的地的评论

● 具体步骤：三步

● Step 1：挑选情感词

- 词性标注（Part of Speech Tagging, POS Tagging）
- 抽取含有**形容词或副词的短语**，它们对于观点通常起了很好的指示作用（上下文作用）

First Word	Second Word	Third Word (Not Extracted)
1. JJ	NN or NNS	anything
2. RB, RBR, or RBS	JJ	not NN nor NNS
3. JJ	JJ	not NN nor NNS
4. NN or NNS	JJ	not NN nor NNS
5. RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

• RORPOSTagger @ A robust and language independent rule-based sentiment classifier, in using English Penn WSJ Treebank sentiment treebank, it achieves 80% accuracy on a computer Core i7-2630M, tagging speed at 80K words/second on a computer Core i7-2630M.
 • SentiTool @ A free online service, includes a sentiment analysis tool, a sentiment analysis API, and a sentiment analysis SDK.
 • Overview of available taggers @
 • Resources for Studying English Syntax Online @
 • CLAWS @
 • LingPipe @ Commercial Java natural language processing library, includes a POS tagger based on the Stanford Log-linear Part-of-Speech Tagger.
 • Apache OpenNLP @ AL 2.0, includes a POS tagger based on the Stanford Log-linear Part-of-Speech Tagger.
 • CRFTools @ Conditional Random Fields (CRFs) English POS tagger.
 • JTextPro @ A Java-based Text Processing Toolkit.
 • Citrus @ LGPL C++ Hidden Markov Model trigram POS tagger.
 • Ninja-POST @ PHP port of GPOSTL, based on Eric Brill's Complexity-Intelligence, LLC @ Free and Commercial NLP.
 • Part-of-Speech tagging based on Soundex features @
 • FastTag - LGPL Java POS tagger based on Eric Brill's Jposc - LGPL Javascript port of FastTag @
 • Topia TermExtractor - Python implementation of the Stanford Log-linear Part-of-Speech Tagger @
 • Northwestern Morphodrone POS Tagger @
 • Part of speech tagger for Spanish @
 • posTagger - Part-of-speech tagger @ Open-source POS tagger

一些POS Tag工具
http://en.wikipedia.org/wiki/Part-of-speech_tagging

24

基于无监督学习的文档情感分类（续）

● Step 2：评估抽取短语的语义倾向（SO）

- 点对互信息：两个词/短语的统计相关性
 - 直观意义：观察到一个词，预测另一个词出现所需的信息量

$$PMI(word_1, word_2) = \log_2 \left(\frac{P(word_1 \wedge word_2)}{P(word_1)P(word_2)} \right)$$

- 短语的情感倾向：基于该短语与正面参考词“excellent”和负面参考词“poor”的关联度计算

$$SO(phrase) = PMI(phrase, "excellent") - PMI(phrase, "poor")$$

- 短语出现的概率可以通过向搜索引擎输入查询所得到的匹配数目进行计算

25

基于无监督学习的文档情感分类（续）

● Step 3：计算所有短语的平均语义倾向

- 给定一条评论，计算评论中所有抽取短语的平均SO值，如果该平均值为正的，则把该评论分类为值得推荐；否则分类为不值得的推荐。

● 论文中最终的分类准确率（Turney, ACL-02）：

- 汽车 - 84%
- 银行 - 80%
- 电影 - 65.83%
- 旅游目的地 - 70.53%

26

文档层次情感分类的优缺点

● 优点

- 提供了对于实体、主题或事件的普遍观点

● 缺点

- 没有描述人们喜欢或不喜欢的细节，这在实际应用中不太好被直接利用
- 不容易应用到非评论的情形中，例如论坛和博客帖子，因为这些帖子会评价多种实体以及对它们进行比较

27

句子层次的情感分析

Sentence Sentiment Analysis

28

句子层次的情感分析

- 文档层次的情感分类对于大部分的应用来说太粗糙了
- 考虑句子层次
- 句子层次情感分析的绝大部分工作重点在于从新闻文章中识别**主观句子**
 - 分类: 客观的和主观的
 - 使用机器学习的某些形式
 - 比如, 使用句子相似度, 朴素贝叶斯分类器, 以及多个朴素贝叶斯分类器
- 得到的主观性句子, 又可以进行**观点倾向 (正面, 负面或中立) (又称极性) 分类**
 - 所用方法与文档层面类似, 但细节的处理有所差异

29

下一步需要考虑什么？

- 在文档和句子层次的情感分类是有用的, **但是**
 - 仍然不能发现观点持有者喜欢什么和不喜欢什么
- 对于对象的一个负面情感
 - 不能说明观点持有者不喜欢对象的任何方面
- 对于对象的一个正面情感
 - 不能说明观点持有者喜欢对象的所有方面
- **我们需要深入到特征/方面层次！**

30

基于方面/特征的情感分类

Aspect-based Sentiment Classification

31

在深入到特征层次之前

- 讨论一下**观点词或短语** (也称作极性单词, 观点支撑单词, 等). 例如,
 - 正面的: beautiful, wonderful, good, amazing
 - 负面的: bad, poor, terrible, cost someone an arm and a leg (idiom)
- 这些词明显对观点挖掘起作用
- 编制或收集观点词列表的三种方法:
 - 人工方法: 可行, 费时费力, 仅是一次性的工作
 - 基于词典的方法: WordNet, HowNet
 - 基于语料库的方法: 依赖大量语料中的语法或共现模式
- **重要提示:**
 - 一些观点词是上下文独立的 (比如, good)
 - 一些观点词是上下文依赖的 (比如, long)

32

基于特征的观点挖掘和汇总

- 以评论为例
- 目标: 找出评论者 (观点持有者) 喜欢什么、不喜欢什么
 - 产品特征及其观点
- 由于关于特定对象的评论很多, 所以必须生成**观点汇总**
 - 值得期待的是一个结构化的汇总
 - 易于可视化和比较
 - 类似但不同于多文档汇总

33

任务

- 在特征/方面层次 (feature/aspect)
 - 任务1: 识别和抽取被观点持有者(即评论人)评价的对象特征
- 任务2: 判定针对该特征的观点是正面的, 负面的还是中立的
- 任务3: 对特征的同义词作分组, 产生多个评论的基于特征的观点**汇总**



34

特征抽取：频繁特征

- **频繁特征**: 被多个评论者多次提及的特征
- 使用**频繁模式挖掘**
- 为什么使用基于频率的方法?
 - 不同的评论者谈论不同的 (不相关的) 事情
 - 当讨论产品特征时, 他们使用趋向于相同或相似的单词.
 - 它们是主要的特征
- 频繁模式挖掘寻找频繁的短语
- Froogle (Google的对比购物网站) 已经实现了该方法 (无词性约束)

35

特征抽取：不频繁特征

- 如何寻找不频繁的特征?
- 观察点: 相同的观点词可以用于描述不同的特征和对象
 - “The **pictures** are absolutely **amazing**.”
 - “The **software** that comes with it is **amazing**.”
- 使用观点和特征的依赖关系抽取特征

• 频繁的特征

• 不频繁的特征



• 观点词



- 此外, 还有一些使用监督学习方法、半监督学习方法、主题模型及聚类技术

36

情感判定或识别

- 对于每个特征, 识别评论者表达的情感或观点倾向
- 基于句子处理, 同时考虑到
 - 一个句子可能包含多个特征
 - 不同的特征可能具有不同的观点
- 例如 :
 - The battery life and picture quality are great (+), but the screen is small (-).
- 几乎所有的方法都使用**观点词**, 但需要注意:
 - 某些观点词具有上下文独立的倾向, 比如, “great”.
 - 某些观点词具有上下文依赖的倾向, 比如, “long”

37

观点词汇总 (观点聚合)

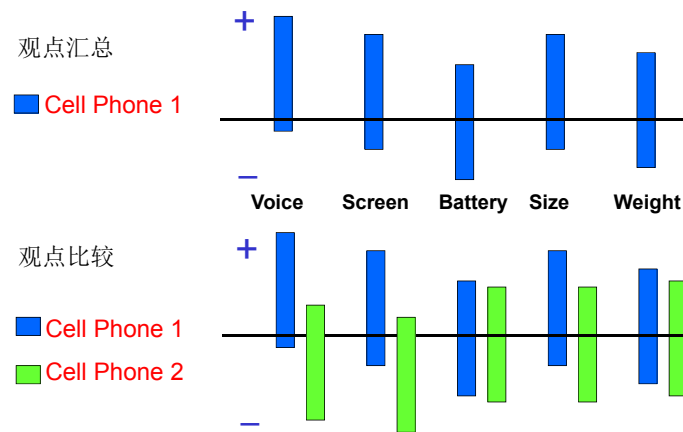
- **输入:** 二元组 (f, s), 其中f是产品特征, s是包含 f 的句子
- **输出:** s中关于 f 的观点的极性
- **两步方法:**
 - 步骤1: 基于转折词 (but, except that等) 切分句子
 - 步骤2: 处理包含 f 的部分s_f. 设s_f中的观点词为w₁, ..., w_n. 对它们的倾向 (1, -1, 0)求总和, 并对(f, s)赋予倾向值
- 在(Ding and Liu, SIGIR-07)中, 步骤2改为

$$\sum_{i=1}^n \frac{w_i \cdot o}{d(w_i, f)}$$

- 得到更好的结果. w_i.o 是w_i的观点倾向值. d(w_i, f)是s中从 f 到 w_i的距离.

38

可视化比较



39

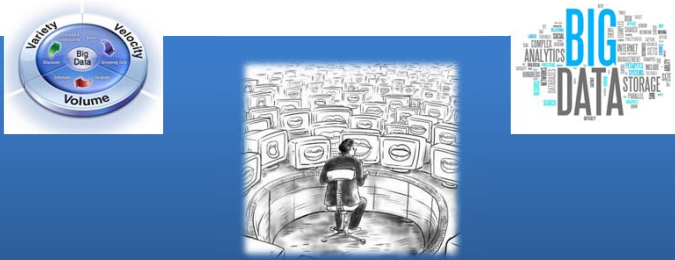
情感分析基础知识：小结

- 情感分析与观点挖掘问题定义
- 情感分析与观点挖掘的任务
 - 文档层次情感分类
 - 句子层次情感分析
 - 基于特征/方面的情感分类
- 总结

40

舆情监测简介

Introductions to Public Opinion Monitoring



41

舆情监测

- 舆情监测是指整合互联网信息采集技术及信息智能处理技术，通过对互联网海量信息**自动抓取**、**自动分类聚类**、**主题检测**、**专题聚焦**，实现用户的网络舆情监测和新闻专题追踪等信息需求，形成简报、报告、图表等分析结果，为客户全面掌握群众思想动态，做出正确舆论引导，提供分析依据。（来源：百度百科）



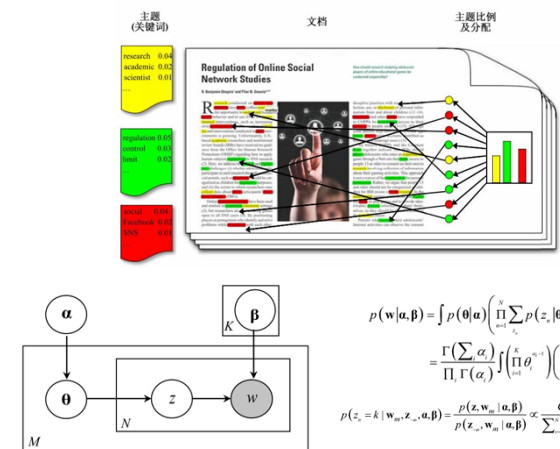
42

公众意见反馈的主题识别及演进分析

- 关键、核心技术：主题建模技术
 - 基础模型：潜在狄利克雷分配（**Latent Dirichlet Allocation, LDA**）模型
 - 是一种语义模型，从文本文档中提取潜在语义信息
- LDA模型背后的直观假设很简单，即为**每一个文档都显示了多个主题信息**。
- 该模型是一种**无监督的生成统计模型**，目标是通过提出一种文档中词语生成的随机过程，找到文档的主题信息。
- LDA模型中假设所有的**主题**是在任何文档生成之前就已经确定。给定语料库中的任何文档，它的生成过程包含两个阶段。首先，随机选择符合狄利克雷随机分布的一个主题分布向量，用以确定该文档中的哪些主题最可能出现。然后，对于文档中出现的每一个词语，随机选择主题分布向量中的一个单独主题。

43

LDA模型



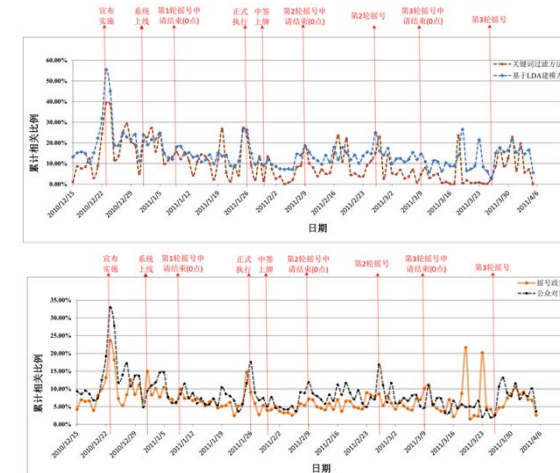
44

LDA主题建模的结果

- **LDA模型可以对所有的文本向量进行训练和推理，然后提取出蕴涵在这些文本数据中的潜在主题信息。通过概率主题建模，可以得到以下两组有用的结果：**
 - **主题关键词列表：**每个潜在主题下最相关的一些词语；
 - **文档-主题概率矩阵：**即每一行对应一个文档，每一列对应一个潜在主题，矩阵中的数值表示对应文档属于对应主题的概率值。
- **潜在主题意义的人工归纳**
- **文档的主题划分/分类**
- **给定时段的特定主题文档数量或累加隶属概率（主题信息量）**

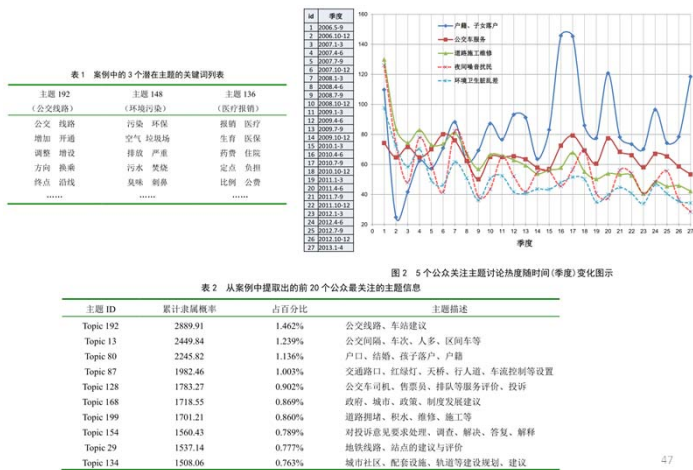
45

案例1：北京小汽车车牌摇号政策反馈分析



46

案例2：北京市某市民公众意见反馈网络平台



47

期末课程论文说明

- **主题要求**
 - 必须与“大数据管理”相关
 - 建议围绕所学专业背景下的“大数据管理问题”展开
- **内容要求**
 - **不少于4000字**，版式：word中正文小四字体，1.5倍行距
 - 独立完成，不得大段拷贝或直接引用网上、书上及他人已发布内容，需要适当引用时请在引用位置注明参考文献来源（**查重**）
 - 论文内容框架（建议）：
 1. 学习本课程的心得体会、感受，对本课程教学的建议和意见（**必有**）
 2. 论文背景介绍
 3. 论文涉及的大数据问题及管理需求、策略和意义（可举实例说明）
 4. 本人对该大数据问题的看法、观点及讨论
 5. 总结
 6. 参考文献和资料

48

期末课程论文说明（续）

● 论文提交要求

- 需要以电子版提交，建议提交word版本
- 作业提交邮箱：bigdata_homework@163.com
- 作业提交截止时间：**第19周周日（2015.01.11）24时**

● 其他说明

- **邮件标题和电子版论文文件请务必按照“学号_班级_姓名.docx”命名，例如“2014211234_2014212103_张三.docx”，也请在邮件中留下姓名、学号及联系方式，以备论文有问题时能够联系到；**
- 请在截止时间之前提交论文（不要在截止时间附近，以避免系统原因过期），过期将不再接收论文提交，成绩为0，请务必注意；
- 每次提交论文后，作业邮箱都会有“已收到邮件”的自动回复，如未收到自动回复，表示发送不成功，请在截止时间内重新提交；
- 论文评分的关注重点
 - 有效的课程建议和意见
 - 关注问题的新颖度
 - 个人分析和讨论的深度
 - 论文的整体工作量

49