中国人民大学
RENMIN UNIVERSITY OF CHINA

多品牌文本评论联动主题模型

Multi-Brand Jointly Latent Dirichlet Allocation Model

CONTENTS

一、研究背景

研究背景

- **评论规模大、增长快**
- **评论挖掘意义重大**

  消费者：真实全面了解产品

  商家：可靠及时的产品改进建议
- **评论挖掘分类**

  语句级挖掘：对产品的整体态度

  特征级挖掘：对产品不同方面特点的态度
- **评论特征挖掘**

  提取评论特征及相应观点

这是一款**氨基酸**洗面奶，我和妈妈一直在用，每次活动时候买，**很划算**，这款洗面奶挤出来就是**泡沫，很细腻丰富**的，**洗脸非常舒服**，不刺激皮肤，**洗的还是很干净的**！

——多芬洗面奶评论（京东）

特征	态度
价格	划算
泡沫	细腻、丰富
肤感	舒服、干净

特征提取方法

1. 传统方法

分类	方法	缺陷
有监督	1. 基于规则抽取 2. 基于机器学习抽取（HMM SVM等） 3. 基于深度学习抽取（命名实体识别、关系提取）	高度依赖打标数据
无监督	1. 基于词频提取 2. Bootstrap提取法 3. 基于语法规则提取	无法合并近义词、同义词

特征提取方法

2. 主题模型方法

效果	代表方法	缺陷
特征提取	Multi-grain LDA(2003)	特征词、态度词混杂
特征 & 情感极性	JST Model(2009) Dependency-Sentiment-LDA(2010) Topic Sentiment Mixture Model(2007)	无法准确提取态度词
特征 & 态度	Local LDA(2010) MaxEnt-LDA(2010) Factorized LDA(2013) Joint Aspect-Based Sentiment Topic(2016)	

本研究出发点

1. 领域内用户普遍关心的特征是什么?
2. 各个品牌内用户关心的特有特征是什么?
3. 用户对这些特征持什么看法?

各品牌的优缺点分别是什么?
其竞争关系是怎样的?

本研究创新点

提出了多品牌文本评论联动主题模型，将评论特性挖掘主题模型推广到了多品牌混合领域。

该模型具有以下优点:

1. 同时提取公共特征、特有特征，了解领域的同时挖掘各个品牌特点。
2. 该模型能够提取各品牌商品在同一特性下的不同态度，便于开展竞争分析。
3. 该模型实现了多品牌语料联合分析，充分利用语料。

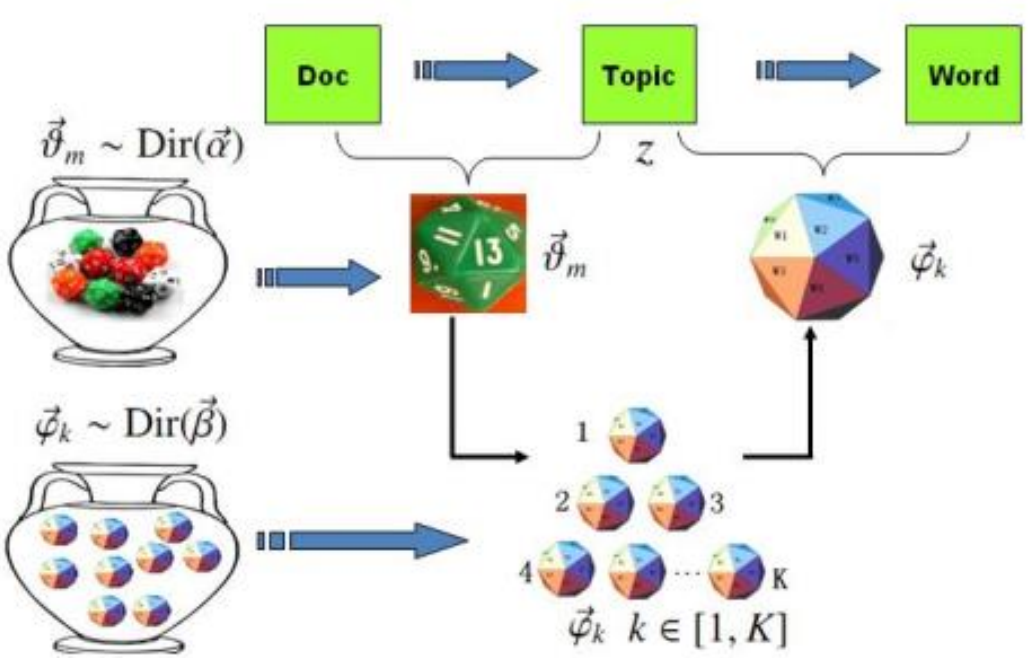
二、模型介绍

理论基础：LDA（2003）

- **一篇文档的生成过程**

1. 上帝有两大坛子骰子，第一坛子装的是doc-topic骰子，第二个坛子装的是topic-word骰子；
2. 上帝随机地从第二个坛子中独立地抽取了K个topic-word骰子，编号为1到K；
3. 每次生成一篇新的文档前，先从第一个坛子中随机抽取一个doc-topic骰子，然后重复如下过程生成文档中的词

- 投掷这个doc-topic骰子，得到一个topic编号
- 选择K个topic-word骰子中编号为z的那个，投掷这个骰子，于是得到一个词



- **求解算法：Gibbs；变分贝叶斯**

理论基础：MaxEnt-LDA（2010）

同时提取：产品整体态度（公共态度）+产品特征&态度

- **两个设计**
  1. 词语分类： 背景词汇、公共特征、特有特征、公共态度、特有态度
  2. 最大熵模型： （背景词、特性词、态度词）先验
- **三个缺陷：无法处理多品牌评论**
  1. 未区分公共主题
  2. 无法提取多品牌对同一特征的不同态度
  3. 无法提取品牌独有特征

步骤 1：生成词汇分布

1. 背景词汇多项分布$\phi^{B} \sim Dirichlet(\beta)$
2. 公共特性词汇多项分布$\phi^{A,g} \sim Dirichlet(\beta)$
   公共态度词汇多项分布$\phi^{O,g} \sim Dirichlet(\beta)$
3. 对特有主题 i $i \in \{1,2 \ldots K\}$
   a. 特有特性词汇多项分布$\phi_i^{A,t} \sim Dirichlet(\beta)$
   b. 特有态度词汇多项分布$\phi_i^{O,t} \sim Dirichlet(\beta)$

步骤 2：生成主题分布、文本

对文章d $d \in \{1,2, \ldots D\}$

1. 主题分布$\theta^{d} \sim Dirichlet(\alpha)$
2. 对句子 s $s \in \{1,2, \ldots S_d\}$
   a. 主题$z_{d,s} \sim Multi(\theta^{d})$
   b. 对句子 s 中词w_{dsn}

$y_{dsn} \sim Multi(\pi_{dsn}), u_{dsn} \sim Bernoulli(p), p \sim Beta(\gamma)$

若$y_{dsn} = 0$（背景词汇），则$w_{dsn} \sim Multi(\phi^{B})$

若$y_{dsn} = 1$，$u_{d_b sn} = 0$（公共特性），则$w_{dsn} \sim Multi(\phi^{A,g})$

若$y_{dsn} = 1$，$u_{d_b sn} = 1$（特有特性）则$w_{dsn} \sim Multi(\phi^{A,z_{d,s}})$

若$y_{dsn} = 2$，$u_{d_b sn} = 0$（公共态度）,则$w_{dsn} \sim Multi(\phi^{O,g})$

若$y_{dsn} = 2$，$u_{d_b sn} = 1$（特有态度），则$w_{dsn} \sim Multi(\phi^{O,z_{d,s}})$

多品牌文本评论联动主题模型

- **三个改进**
  1. 增加公共主题层次

     可提取不同品牌在同一公共特征下不同表现
  2. “品牌联动”
     a. 公共主题在大词库上更新
     b. 特有主题在品牌独有小词库上更新
  3. 允许各品牌特有主题数不同

- **模型求解：Gibbs 采样**

步骤 1：生成词汇分布

1. 背景词汇多项分布$\phi^{B} \sim Dirichlet(\beta)$
2. 对公共主题 i $\in \{1,2,3..K\}$
   公共特性词汇多项分布$\phi_{i}^{GA} \sim Dirichlet(\beta)$
3. 对品牌 b 公共主题i i $\in \{1,2,3..K\}$ b $\in \{1,2,3..B\}$
   公共态度词汇多项分布$\phi_{ib}^{GO} \sim Dirichlet(\beta_b)$
4. 对品牌 b 特有主题 j b $\in \{1,2 \dots B\}$ j $\in \{1,2 \dots K_b\}$
   a. 特有特性词汇多项分布$\phi_{ib}^{SA} \sim Dirichlet(\beta_b)$
   b. 特有态度词汇多项分布$\phi_{ib}^{AO} \sim Dirichlet(\beta_b)$

步骤 2：生成主题分布、文本

对品牌 b 文章d d $\in \{1,2, \dots D_b\}$

1. 公共主题分布$\theta_{d}^{G} \sim Dirichlet(\alpha_g)$
2. 特有主题分布$\theta_{d}^{S} \sim Dirichlet(\alpha_s)$
3. 对句子 s s $\in \{1,2, \dots S_d\}$
   a. 公共主题$z_{d_b s}^{G} \sim Multi(\theta_d^G)$，特有主题$z_{d_b s}^{S} \sim Multi(\theta_d^S)$
   b. 对句子 s 中词w_{dsn}

$y_{d_b sn} \sim Multi(\pi_{dsn}), u_{d_b sn} \sim Bernoulli(p), p \sim Beta(\gamma)$

若$y_{d_b sn} = 0$（背景词汇），则$w_{d_b sn} \sim Multi(\phi^{B})$

若$y_{d_b sn} = 1$，$u_{d_b sn} = 0$（公共特性），则$w_{d_b sn} \sim Multi(\phi_{z_d^g}^{GA})$

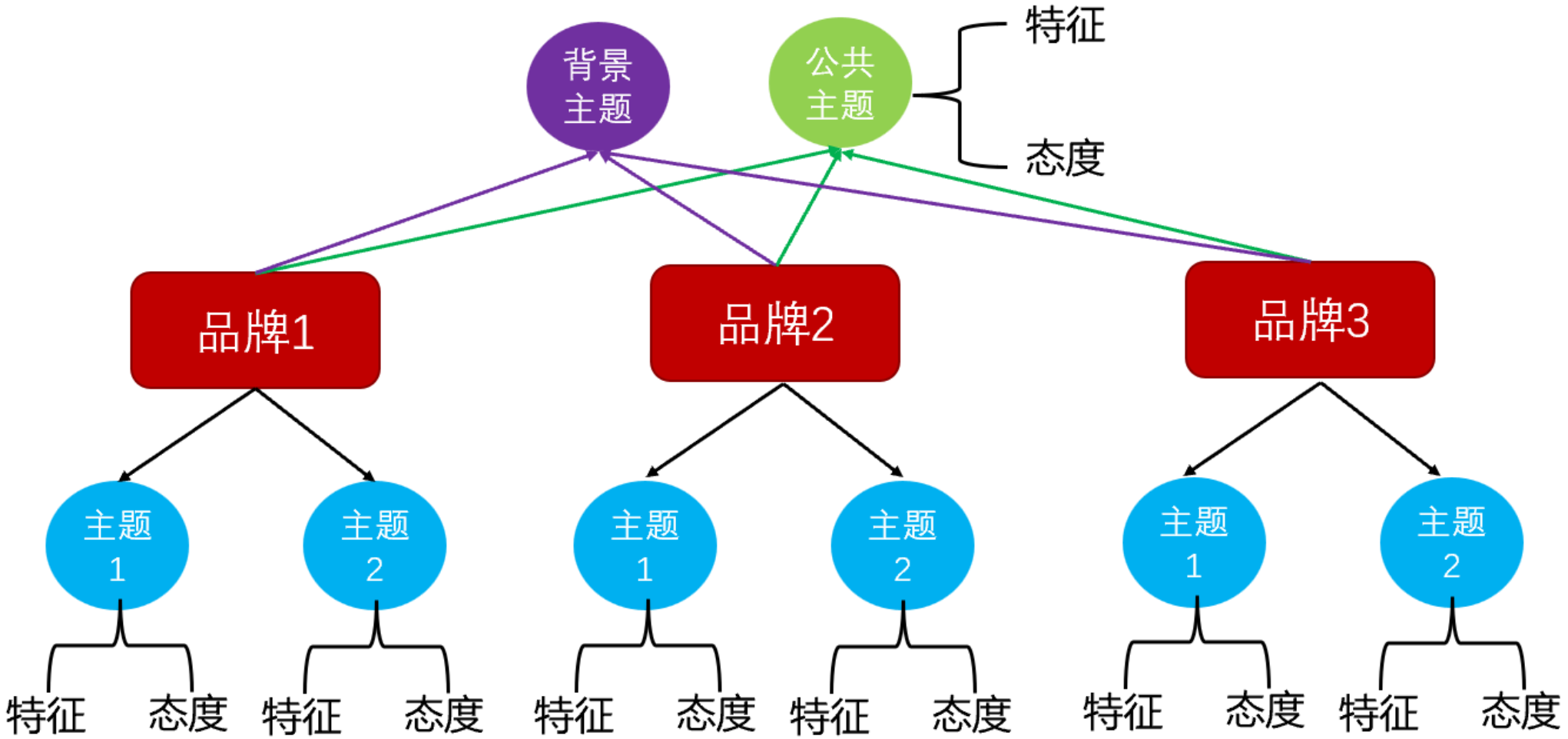
若$y_{d_b sn} = 1$，$u_{d_b sn} = 1$（特有特性）则$w_{d_b sn} \sim Multi(\phi_{z_d^s b}^{SA})$

若$y_{d_b sn} = 2$，$u_{d_b sn} = 0$（公共态度）,则$w_{d_b sn} \sim Multi(\phi_{z_d^g b}^{GO})$

若$y_{d_b sn} = 2$，$u_{d_b sn} = 1$（特有态度），则$w_{d_b sn} \sim Multi(\phi_{z_d^s b}^{SO})$

多品牌文本评论联动主题模型

目标：同时提取产品整体态度（公共态度）+产品特征&态度



三、实证分析

护肤品数据：数据介绍&预处理

2018年12月— 2019年2月　京东商城洗面奶评论

品牌	价格	产地	评论数	评论平均长度
旁氏	30	中国	499	37.83
多芬	45	日本	643	31.28
珂润	108	日本	520	25.78
丝塔芙	109	加拿大	1170	44.60
娇韵诗	250	法国	545	30.07
雪花秀	320	韩国	561	27.44

文本预处理&词云图

预处理

1. 数据清洗&分词
2. 词语类别标注
   a. 特征词——名词 / 人工标注特性集合
   b. 态度词——HOWNET / NTUSD
   c. 背景词
3. 概率标注（最大熵模型）

多芬品牌语料词云图

护肤品数据：主题数选择

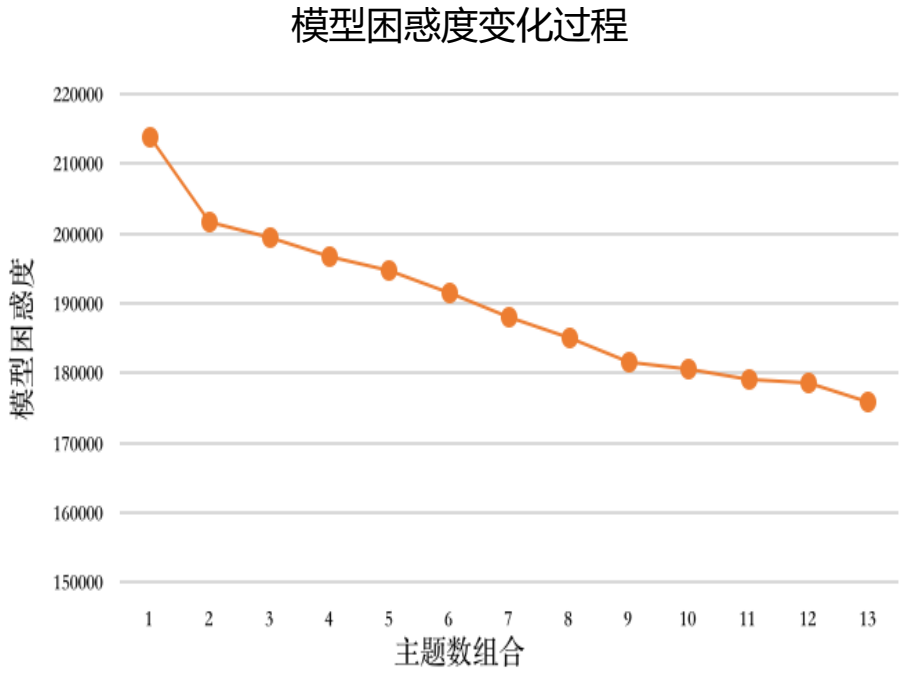
- **待调参数多**
  1. 公共主题数
  2. 各品牌特有主题数

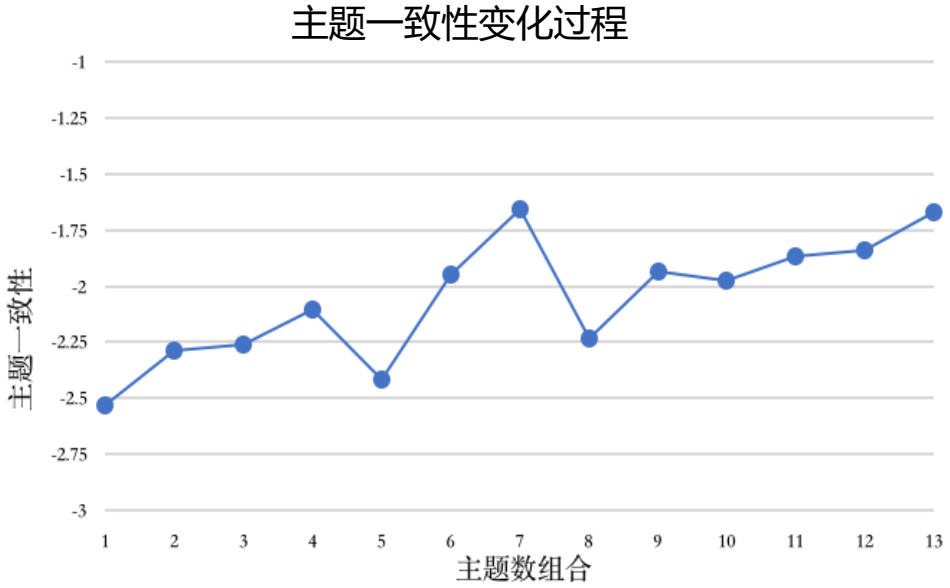
- **前向逐步选择法**

  从空模型开始，逐步增加公共主题数及各品牌主题数使模型困惑度下降最多的模型

- **最优主题数组合**

  公共主题数：2
  特有主题数：多芬3，娇韵诗2，柯润3，旁氏4，丝塔芙3，雪花秀3





护肤品数据：模型结果

公共主题1

特征		成分 **神券** 洗过 **验证** 太久 **评价** **真假** **凑单** 状况 护肤用品	→ **售卖店铺 / 平台**
态度	多芬	反馈 油敏皮 **争议** 态度 给力 **漏液** 速度 **挂羊头卖狗肉**	→ **争议、漏液**
	娇韵诗	**贵** 真心 缺点 不愧 很早 **OK** **快** **大赞** 可用 浮动	→ **贵、OK、大赞**
	珂润	**合适** 试用 软软 有待 还行 **划算** 备用 快 卖 **太棒了**	**划算、好用**
	旁氏	送货 淡淡 化学 **物美价廉** 回去 特意 加油 改天 **使用方便**	
	丝塔芙	**过敏** **差远了** **舒服** 两遍 护肤品 有意思 **起痘** 积分 惊喜 **闷痘**	**褒贬不一**
	雪花秀	**失望** 打开 入手 滑滑 清 粉 **挺舒服** 包装盒 卸妆乳 沐浴	

护肤品数据：模型结果

多芬：特有主题

主题1	特征 **物流** **包装** 仔细 很好 很快 满意 完全一致 **卖家** 第一次 态度 **喜欢** 购买 **不错** 替换 **好评** 物流 很快 超出 期望值 支持	**物流包装不错**
主题2	特征 **泡沫** 感觉 **敏感肌** **正品** 瓶装 **速度** 价格 保湿 **成分** **质量** 态度 **好用** **不错** 快 很好 **温和** 很多 **喜欢** **适合** 刺激 会回购	**泡沫温和、好用**
主题3	特征 **效果** 想买 **油性皮肤** 水乳 挺 油腻 去年 **控油** 一套 态度 **不错** **很好** 特别 油腻 用过 干净 希望 好用 **舒服** 控油	**适合油性皮肤、控油**

护肤品数据：模型结果

丝塔芙：特有主题

主题1	特征 丝塔芙 **物流** 没 **速度** 干净 **很快** 用过 快用 好用 适合 态度 舒服 快 **物流** 丝塔芙 **给力** 服务 购买 信赖 包装	**物流很快**
主题2	特征 **泡沫** **效果** 丝塔芙 价格 量 紧绷 **敏感肌** 黑头 脸上 态度 好用 不错 没有 **刺激** **干净** 习惯 **紧绷** 喜欢 很好	**敏感肌适用**
主题3	特征 客服 套装 收到 没想到 商品 包装盒 **打开** 小瓶 **糊状** 态度 包装 感觉 生产 泡沫 价格 商品 **不好** 瓶子 **退货**	**质地不好**

护肤品数据：结果分析

- **相同点**

  温和、泡沫丰富、清洗干净、物流快

- **不同点**

1. **适用人群不同**
   a. 多芬、丝塔芙、旁氏洗面奶适合敏感肌
   b. 娇韵诗洗面奶适合怀孕人群
   c. 多芬适用于油性肌
2. **成分不同**
   a. 旁氏、雪花秀为氨基酸洗面奶，
   b. 其余为普通洗面奶
3. **营销策略不同**
   a. 娇韵诗主打孕期可用
   b. 旁氏主打米粹
   c. 雪花秀主打其独特香味
   d. 旁氏、珂润赠品丰富
4. **客服态度不同**
   a. 丝塔芙、珂润、多芬的客服态度较差

护肤品数据：后续分析

- **思路**

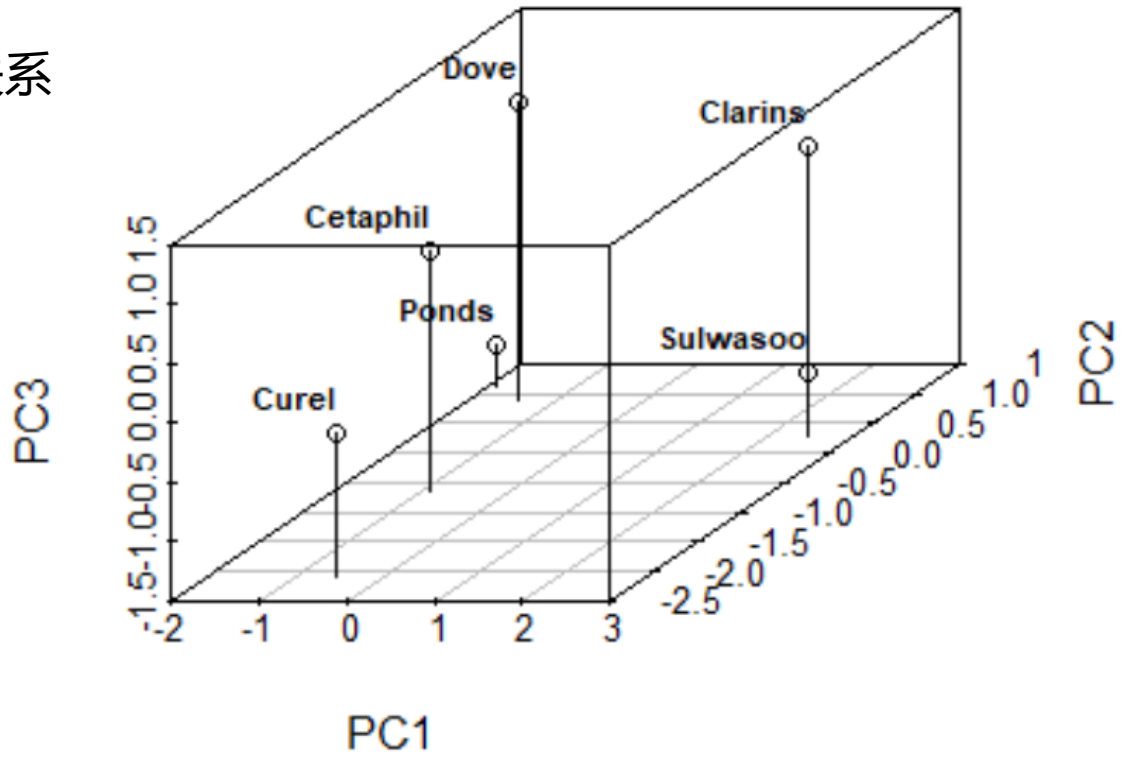
  利用提取出的商品特征，分析品牌之间的竞争关系

- **方法**

  整合公共特征和特有特征得到特征集合
  计算各个品牌在每个特征上的Jaccard系数

  Jaccard=a/(F_1+F_2 −a)
  a : 各个品牌中包含指定特征的文档数
  F_1: 各个品牌的文档数
  F_2: 所有品牌的文档数

  Jaccard系数用以衡量特征对品牌的重要程度
  对六个品牌的Jaccard系数进行PCA分析



日本菜餐厅评论：数据介绍

2006年4月—2020年6月 大众点评“将太无二”连锁评论

	人均消费	总评分	评论数
朝阳大悦城店	155	4.65	11280
丰台万达广场店	147	4.70	1413
恒泰广场店	138	4.69	1741
欧美汇购物中心店	144	4.78	14992
瑞士公寓餐厅店	126	4.67	9860
西单汉光店	145	4.66	3389

欧美汇分店词云图

日本菜餐厅评论：模型结果

公共主题6

特征		拉面 面 地狱 骨汤 汤		→ 拉面类菜品
		好评	差评	
态度	朝阳大悦城	清淡 茶泡饭 **不推荐** 微微 挺辣	**糟糕 难吃 不推荐** 辣 足够	} 一致不推荐
	丰台万达广场	便宜 **难喝 无敌 难吃** 适宜	**逊色 难吃 下降 糟糕** 分熟	
	恒泰广场	**好吃** 可爱 **辛辣** 顺利 解决	常买 **没盐** 记住 熬制 宽	
	欧美汇购物中心	牛仔 拉面 地狱 鱿鱼 **不吃**	**很辣** 凉 湘菜 **香 浓重**	
	瑞士公寓餐厅	**过瘾** 恢复 松茸 **新颖** taste	**很辣** 炒作 不加 **香 鲜**	→ 一致推荐
	西单汉光	**挺好 不辣** 奇怪 **浓稠 浓厚**	**平平** 没给 **便宜 浓郁** 煮开	

日本菜餐厅评论：模型结果

		公共主题3		
	特征	**环境 装修 服务 位置 风格**		→ **环境**
		好评	差评	
态度	朝阳大悦城	适合 优雅 舒服 舒适 干净	宽敞 干净 舒适 破旧 拥挤	
	丰台万达广场	优雅 五层 经营 靠边 舒适	不爽 贵 拥挤 宽敞 拉低	
	恒泰广场	**简洁 舒服 明亮** 私密 独特	开心 嘈杂 站 **妥妥 明亮**	→ **环境宽敞明亮**
	欧美汇购物中心	靠窗 适合 中关村 安静 舒服	嘈杂 局促 说话 管理 未变	
	瑞士公寓餐厅	**适合 拥挤 卫生 难 昏暗**	**拥挤 冷 局促 压抑 昏暗**	} **环境拥挤嘈杂**
	西单汉光	**嘈杂 拥挤 放不下 摆不下 差**	**拥挤 矮 昏暗 难受 差**	

日本菜餐厅评论：结果分析

- **公共推荐菜**

  寿司、三文鱼、秋天的童话、北极贝、海胆等

- **特有推荐菜**

  a) 瑞士公寓：拉面
  b) 朝阳大悦城、丰台万达广场、恒泰广场：鳗鱼、牛油果、沙拉、天妇罗等

- **褒贬不一菜品**

  a) 牛肉、火锅：清淡可口 / 很咸。
  b) 恒泰广场、欧美汇购物中心、西单汉光的拉面：过瘾 / 太辣
  c) 欧美汇、瑞士公寓、西单汉光的鳗鱼、牛油果、沙拉、天妇罗：很嫩 / 粗糙。

- **一致不推荐菜品**

  朝阳大悦城、丰台万达广场：拉面

- **环境**

  瑞士公寓餐厅、西单汉光：环境差、局促拥挤

四、未来展望

后续工作

- **结合稀疏主题模型**

  考虑主题稀疏或词稀疏，有助于增大主题区分度，增强可解释性

- **结合动态主题模型**

  考虑时间变化，挖掘品牌的主题演化

- **结合词向量模型**

  借鉴词向量思想，获得特征向量协助下一步商业分析

感谢聆听