

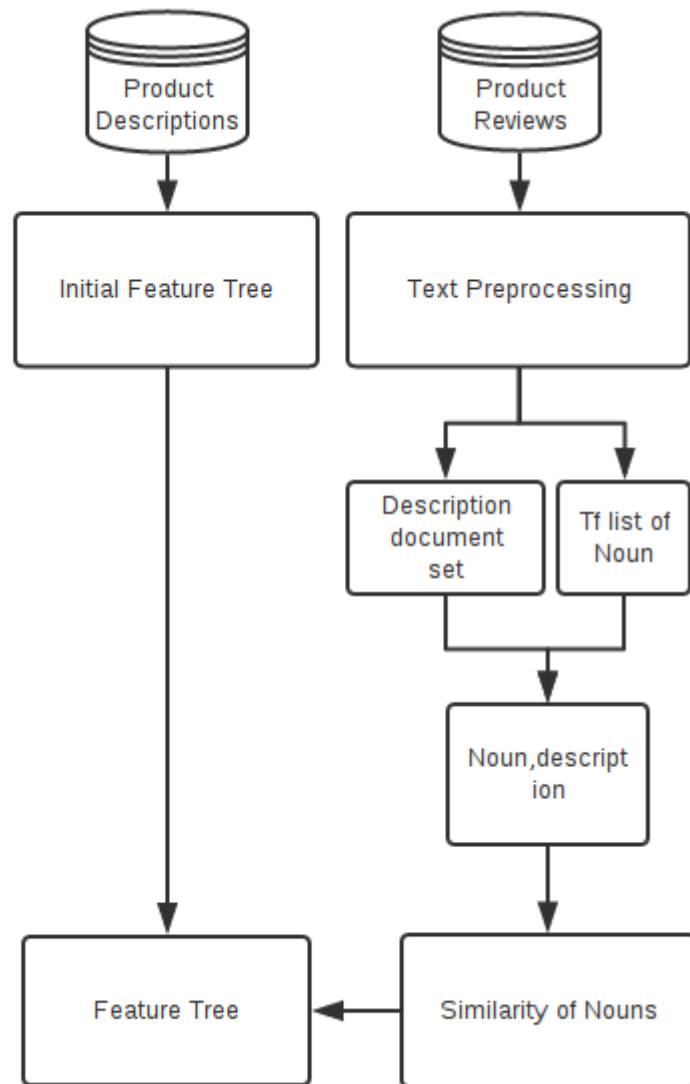
数据来源与方法框架

一，叙论：

近年来，电子商务行业发展迅速，业类各大公司都积累了大量的数据，其中包括顾客的商品评论。同时，有大量的研究指向商品评论。[关于商品评论研究的文献回顾](#)。[有关于语义挖掘的文献综述](#)。

在这里我们给出来一种方法，对评论文本内容进行概念建模。生成了顾客在给出评价时所依照的概念树。

研究框架如下：



二，数据来源：

在线商品与商品描述（eg，与手机有关）。

1，京东商城商品，描述，评论。

★收藏京东

的哈，欢迎来到京东！[退出] | 我的订单 | **VIP** 会员俱乐部 | **B2B** 企业频道 | 手机京东 | 客户服务 | 网站导航



170特权卡

搜索

热门搜索：还有机惠 锤子手机 OPPO N3 中兴 V5S 免费流量 JD Phone配件

我的京东

去购物车结算

全部商品分类

首页

服装城

食品

团购

寻宝岛

闪购

金融

手机 > 手机通讯 > 手机 > 苹果 (Apple) > 苹果iPhone 5s



苹果 (APPLE) iPhone 5S 16G版 4G手机 (金色) FDD-LTE/TD-LTE/TD-SCDMA/WCDMA/GSM 兼容移动/联通网络iPhone5s (A1530)，畅享4G/3G高速网络！（购买前请详阅以下温馨提示）【买卡上京东，省钱又轻松】点击有惊喜！

京东价：**¥4199.00** (降价通知)

🔥 上涨

💰 更低价 ¥4099

★ 收藏

商品编号：1057746

促销信息：**🎁 赠下方1件热销商品，赠完即止** (共2项促销) 查看

 鑫盾 高清透越耐磨手机贴膜 适用于苹果iPhone5/5S 前后 新... × 1

商品评分：**★★★★★** (已有28692人评价) 

配送至：

四川成都市高新西区

有货，23:00前完成下单，预计明日 (11月25日) 送达

商品介绍

规格参数

包装清单

商品评价 (28692)

售后保障

加入购物车

主体

品牌

苹果 (Apple)

型号

iPhone 5s

颜色

金色

上市年份

2014年

输入方式

触控

智能机

是

3G视频通话

需使用第三方软件

操作系统

苹果 (IOS)

操作系统版本

IOS

CPU品牌

苹果

CPU型号

A7芯片 M7运动协处理器

CPU频率

1.3GHz

CPU核数

双核

运营商标志或内容

无

网络

4G网络制式

移动4G (TD-LTE) /联通4G (FDD-LTE) /联通4G (TD-LTE)

3G网络制式

移动3G(TD-SCDMA)/联通3G(WCDMA)

2G网络制式

移动2G/联通2G(GSM)

网络频率

UMTS (WCDMA)/HSPA+DC-HSDPA (850、900、1900、2100 MHz)
GSM/EDGE (850、900、1800、1900 MHz)
TD-SCDMA 1900 (F)、2000 (A)
TD-LTE (频段 38、39、40)

数据业务

802.11a/b/g/n 无线网络 (802.11n 时工作在 2.4GHz 和 5GHz)

存储

2, 亚马逊网站的商品, 商品描述

amazon

Try Prime

Search

Unlocked Cell Phones +

Go

Your Account ▾

Departments ▾

Fire & Kindle ▾

Recommended for You

Today's Deals

Gift Cards

Help

Sell

Cell Phones & Accessories

Contract Phones

No-Contract Phones

Unlocked Phones

Accessories

Cases

Best Sellers

Deals

Trade-In

All Electronics

ELECTRONICS

HOLIDAY GIFT GUIDE

Great Deals Every Day
→ Shop now

Show results for

< Cell Phones & Accessories

< Cell Phones

Unlocked Cell Phones

Refine by

Featured Brands

☐ Samsung (1,219)

☐ BLU (473)

☐ blu (473)

☐ Motorola (455)

☐ BlackBerry (437)

☐ Apple (528)

☐ Nokia (768)

+ See more

Display Size

☐ 3.9 Inches & Under (4,397)

Unlocked Cell Phones

Eligible for Free Shipping

Free Shipping by Amazon

Color

☐ Black

☐ White

☐ Silver

☐ Blue

☐ Yellow

☐ Red

☐ Pink

☐ Green

☐ Gold

Operating System

☐ Android (8,478)

☐ BlackBerry (497)

☐ iOS (759)

☐ Windows Phone (553)

Phone Type

☐ Touchscreen (8,090)

☐ Smart Phone (7,234)

☐ Bar (1,708)

☐ Basic Phone (1,361)

☐ Flip (775)

☐ Slider (217)

Internal Storage

☐ 4 GB (2,078)

☐ 8 GB (1,944)

☐ 16 GB (2,259)

☐ 32 GB (1,641)

☐ 64 GB (165)

Feature Keywords

☐ GSM (11,376)

☐ Dual SIM (5,207)

☐ Quad Band (2,808)

☐ QWERTY Keyboard (1,275)

live in Color

The All New BLU C Series.

>Learn more

Blackberry Passport

Work Wide

Live in Color

The all-new BLU C

Built to Inspire E

HTC One (M8)

Splash of Color

Unlocked iPhone 5

BLU JR Unlocked

Windows Phone

Learn more

3，淘宝网的商品，描述。

淘宝数码
3c.taobao.com

更多市场

宝贝

首页手机相机单反PC电脑智能3C馆智能家电

所有分类 > 手机

共 5859 款产品 查看全部相关宝贝

品牌:	小米	苹果	三星	华为	酷派	诺基亚	HTC	联想	多选	更多
	索尼	vivo	OPPO	魅族	大显	中兴	福中福	LG		
尺寸:	3.0英寸以下	3.0-3.9英寸	4.0-4.5英寸	4.6-4.9英寸	5.0-5.5英寸	5.6-6.0英寸	多选	更多		
操作系统:	Android/安卓	iOS	无操作系统	Windows phone	BlackBerry/黑莓	阿里手机系统	多选	更多		
网络类型:	联通4G(FDD-LTE)	移动4G(TD-LTE)	电信4G(FDD-LTE)	联通3G(WCDMA)	移动3G(TD-SCDMA)	多选	更多			
筛选条件:	像素	核心数	分辨率	运行内存RAM	上市时间					

4，ebay网商品，描述

Hill Sign in or register | Daily Deals | Gift Cards | Sell | Help & Contact

Shop by category

iphone 5

Related: iphone 4s iphone 5 unlocked samsung galaxy s4 iphone 5 verizon samsung galaxy s

Categories

Cell Phones & Accessories

Cell Phones & Smartphones

Features

3G Data Capable (9,739)

4G Data Capable (9,808)

Bluetooth Enabled (12,349)

GPS (11,047)

Music Player (11,611)

Price

RMB to RMB

Format

All Listings (31,054)

Auction (16,972)

Buy It Now (18,877)

Delivery Options

Free international shipping

Show only

Returns accepted

Completed listings

Sold listings

More refinements...

Related Buying Guides

Apple iPhone 5 vs. Samsung Galaxy S4

How to Use the iPhone 5

All Listings Auction Buy It Now

Sort:

31,054 results for iphone 5 Follow this search

iPhone 5

Carrier

All Carriers

About Unlocked Phones

Storage Capacity

16 GB 32 GB

64 GB

Clear all refinements

Features

Model

Carrier

Color

Storage Capacity

Brand

Operating System

Style

Contract

Condition

Price

Format

Seller

Delivery Options

3G Data Capable (9,735)

4G Data Capable (9,805)

Air Gesture (20)

Bluetooth Enabled (12,347)

Camera (647)

Email Access (1)

Fingerprint Sensor (6,537)

Global Ready (227)

GPS (11,044)

Internet Browser (10,979)

Iris Scanning (15)

Music Player (11,609)

Near Field Communication (23)

Push to Talk (102)

QWERTY Keyboard (171)

Speakerphone (12,017)

Touchscreen (11,201)

TTY Compatible (184)

Voice-Activated Dialing (315)

Wi-Fi Capable (12,341)

Wireless Charging (24)

(0) Filters selected

三，生成概念树

以京东商城 IPHONE5S 商品为例。包括其商品描述和商品评论。

1,基于商品描述构造初始概念树。

根节点：京东

第一层：手机

第二层：品牌 - 价格 - 网络 - 系统 - 机身，颜色 - 显示 - 主体 - 存储 - 感应器 - 摄像 - 娱乐 - 传输 - 其它

第三层：品牌：华为 - 苹果 - 三星 - 酷派 - 努比亚 - htc-诺基亚 - 魅族 - 联想 - 索尼 - 中兴。。。

价格：

网络：移动 - 电信 - 联通 - 2G-3G-4G-双卡-网络-制式-数据-业务-频率

系统：安卓-苹果-微软-IOS-windowsphone

机身，颜色：白色-灰色 - 黑色 - 金色 - 银色 - 红色 - 蓝色 - 粉色 - 黄色 - 绿色 - 橙色 - 紫色

显示：屏幕-材质-尺寸-触摸屏-分辨率-字体-色彩

主体：型号-输入-智能机-视频通话-版本-CPU-频率-核数

存储：内存-运行-存储卡-类型-16G-ROM

感应器：gps-重力-光线-距离

摄像：摄像头-闪光灯-对焦

娱乐：收音机-音乐-视频-电视

传输：WIFI-蓝牙-热点

其它：SIM 卡-尺寸-电池-型号-类型-通话-时间-待机-数据线-耳机-接口-尺寸-重量

讨论点（1，每一个结点的名词也是该结点的特征。而结点“其它”并不是一个有效的名词特征。在评论文本中，“其它”作为一个词很少出现，而位于“其它”下的特征名词则能作为该类的特征，因而，对于结点“其它”并不会对概念树的生成造成恶劣的影响。2,每个结点可能是一个词，也可能是 N 个词。）

在实验中，我们以根节点到第二层树的生成为例。如果构造的初始概念树有 N 层，我们的方法也可以用于生成 N 层的概念树。

2,计算数据集中名词与名词之间的相似性。

按句子隔断后一般结果构成（目标名词+其他名词 + 动词 + 形容词+其他词）。

句子的主题是名词，其他词性作为名词的描述词出现。将每一个名词的描述词集合作为一个文档，训练 TopicModel。根据每个文档的 topic 分布，利用 KL 距离，度量名词之间的距离。

语义上的假设：同义词往往对应相似的描述词，描述词的差异越大，其语义差距越大，在语义 topic 上的分布，差异越大。

- 文本预处理与名词词频统计

分词（词性标记）-去停用词

预处理并按照句号，问号，感叹号（通常认为的语义切换符号）。对文档进行分段处理。

词频统计	名词词频统计	其他词性词频统计
不错/a 8787	手机/n 4643	不错/a 8787
手机/n 4643	京东/ns 3319	买/v 3890
买/v 3890	苹果/n 3065	不/d 2512
京东/ns 3319	正品/n 2658	喜欢/v 2055

苹果/n 3065	东西/ns 1786	好好/d 1442
正品/n 2658	速度/n 1452	说/v 1318
不/d 2512	5s/n 1254	送货/v 1077
喜欢/v 2055	感觉/n 1252	满意/v 1058
东西/ns 1786	质量/n 986	很快/d 1046
速度/n 1452	系统/n 957	给力/a 965
好好/d 1442	价格/n 803	送/v 869
说/v 1318	土豪/n 746	很好/a 843
5s/n 1254	朋友/n 724	值得/v 825
感觉/n 1252	4g/n 698	很好/d 714
送货/v 1077	行货/n 680	支持/v 701
满意/v 1058	物流/n 664	太/d 693
很快/d 1046	信号/n 565	快递/v 693
质量/n 986	电池/n 537	发现/v 627
给力/a 965	评价/n 533	流畅/a 621
系统/n 957	屏幕/n 510	赞/v 572

- 生成 名词-名词描述文档。
 - 1,文档中不存在名词时，对该句不理睬。
 - 2,句子中不只一个名词。可能的假设：**其他名词是某一个名词的附属词，按词频排序，只计入高频名词的描述文档。**

algorithm

```

Def find_affiliate(word,content) :
    new_content=[word," "]
    content_copy=content[:]
    for line in content:
        if word in line:
            new_content[1]+=line
            content_copy.remove(line)
    return new_content,content_copy
content=[]
Noun=[]
Noun_affiliate=[]
for word in Noun:
    if len(content)==0:
        break
    if word in content:
        new_content,content=find_affiliate(word,content)
return Noun_affiliate[word,word_affiliate]

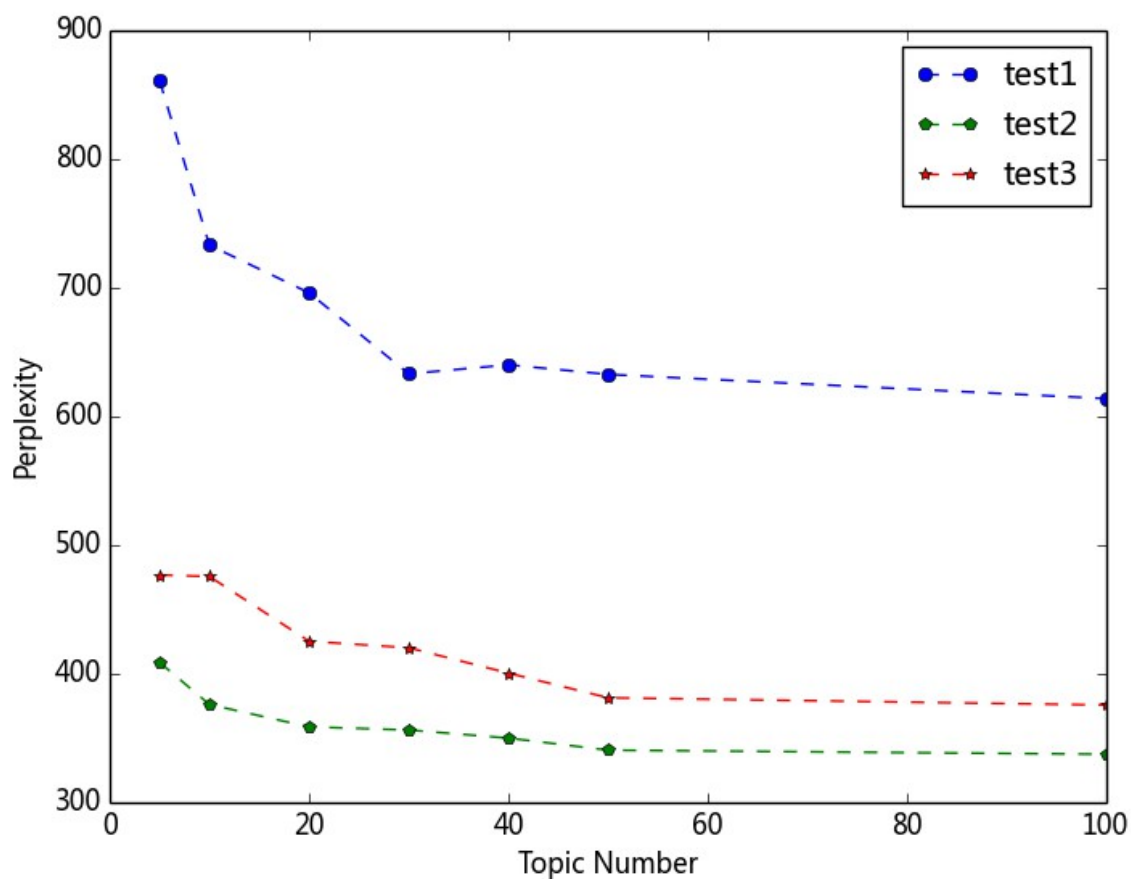
```

#定义一个从分段文档集中获取名词附属词的函数。
#预生成的附属词文档集，列表的第二项为附属词文档集
生成一个分段文档的副本
#循环读取分段文档集中的文档
如果目标名词在该分段文档中
#则该分段文档所有词都是该词的附属文档，包括其本身
在分段文档集的副本中删除该文档
#返回词 - 附属文档以及下一个词的待查找分段文档
#第二步中，根据标点生成的分段文档集合
#名词列表，根据名词排序生成的名词列表
#如果待查找分段文档集长度为空，则终止循环
#如果 word 在 content 中，引用函数，生成该 word 的附属文档集。同时将分段文档集重置
#得到所需要的词-附属文档集合

手机/n	iphone/eng 现时/t 性价比/n 一款/m 手机/n 不错/a 果/ng apple/eng iphone/eng 5s/n 16/m g/eng 版/n 4g/n 手机/n 金色/n fdd/eng lte/eng td/eng lte/eng td/eng scdma/eng wcdma/eng gsm/eng 攒/v 物流/n 当天/t 下单/v 当天/t 到货/v 手机/n 盖/v 两个/m 黑点/n 接受/v 到手/d 三天/m 降价/n 吭/zg 京东/ns 物流/n 值得/v 肯定/v 下单/v 签收/v 时间/n 手机/n 耍/v 两天/m 死机/n 重启/v 电量/n 时间/n 无故/n 掉/v 很快/d 手机/n 很好/d 很漂亮/n 质量/n 不错/a 手机/n 情况/n 不错/a 苹果/n 手机/n 信赖/n 苹果/n 手机/n 没话说/l 系统/n 给力/a 期待已久/i 土豪/n 金/nr 终于/d 到手/d 。。。
机子/n	拿到/v 机子/n 不错/a 发现/v 换/nz sim 卡/n 跑/v 好几个/m 地方/n 差点/n 动手/n 剪/v 还好/v 机子/n 当日/t 激活/a 有时候/l 发热/v 正规/a 发票/n 未/d 拆封/v 机子/n 通用/v 网络/n 正版/v 机子/n 不错/a 机子/n 实用/v 机子/n 还行/v 自动/vn 卡死/nr 自动/vn 跳/v 回主/v 屏/n 反正/d 保修/v 一会/m 机子/n 热/a 暂时/d 发现/v 评论/n 担心/v 运气/n 还好/v 一点/m 查/v 序列号/n 10/m 月/m 20/m 号/m 生产/vn 机子/n 机子/n 新/a 顺畅/a 机子/n 不错/a 发现/v 全新/d 未激活/i 不错/a 新/a 机子/n 14/m 年/m 快递/v 给力/a 。。。
价钱/n	还行/v 听/v 顺手/d 价钱/n 高/a 一分/m 价钱/n 一分货/n 高大/a 价钱/n 降低/v 星期/t 降/v 100/m 块/q 不错/a 奥/nr 掉/v 太/d 价钱/n 不错/a 一分/m 价钱/n 一分货/n 不错/a 价钱/n 优惠/vn 服务到位/n 送货/v 到货/v 很快/d 价钱/n good/eng 不错/a 价钱/n 公道/n 实惠/vn 草/n 显示/v 无卡/ns 花/v 价钱/n 买/v 不/d 舒心/a 一分/m 价钱/n 一分货/n 不错/a 上个月/t 入手/v a1518/eng 4799/m a1530/eng 4799/m 20/m 抵用/vn 券/n 4779./m。。。
价格/n	价格/n 确实/ad 不错/a 商场/n 便宜/a 价格/n 实惠/vn 不错/a 价格/n 合适/a 送/v 贴膜/n 完美/a 送给/v 老婆/n 天/q 暂时/d 发现/v 考察/v 送货/v 很快/d 价格/n 实惠/vn 买/v 第三个/m 价格/n 实惠/vn 性价比/n 高/a 价格/n 合适/a 土豪/n 金/nr 高大/a 价格/n 给力/a 幸好/a 双十/m 价格/n 价格/n 算/v 土鳖/n 买得起/v 快递/v 效率/n 高/a 昨天/t 下单/v 支持/v 买来/v 送/v 价格/n 浮动/vn 很大/a 不错/a 价格/n 算是/v 优惠/vn 第二次/m 京东上/n 买/v 类似/v 价格/n 电子产品/n 还好/v 失望/v 看着/v 评论/n 犹豫/a 很久/m 买/v 到手/d 查/v 。。。
外观/n	外观/n 漂亮/a 唯一/b 音量/n 键/n 松动/a 整体/n 不错/a 网上/s 查/v 新机/n 确实/ad 不错/a 喜欢/v 外观/n 超/v 喜欢/v 通话质量/n 外观/n 漂亮/a 功能齐全/i 通话音质/l 功能齐全/i 外观/n 漂亮/a 值得/v 拥有/v 外观/n 磨损/vn 外观/n 磨损/vn 外观/n 磨损/vn 外观/n 磨损/vn 外观/n 磨损/vn 外观/n 很漂亮/n 好看/v iphone5s/n 外观/n 不错/a 外观/n 漂亮/a 喜欢/v 一惯/n 喜欢/v 外观/n 不/d 解释/v 软件/n 不错/a 外观/n 漂亮/a 待机时间/n 长/a 顺手/d 外观/n 很漂亮/n 分辨率/n 高/a 外观/n 漂亮/a 铃声/n 外观/n 漂亮/a 外观/n 颜色/n 通话/n 清晰/a 外观/n 漂亮/a 操作/v 真心/d 值/n 钱/n 性价比/n 外观/n 不错/a 用起/v 麻烦/an 外观/n 向往/d 值得/v 赞/v 外观/n iphone5/eng 指纹/n 还好/v 不错/a 外观/n 很漂亮/n 。。。
金色/n	帮/v 同事/n 买/v 额/n 两个/m 几个/m 金色/n 黑色/n 不错/a 满意/v 网购/n 金色/n 黑色/n 贵/a 几百块/m 办法/n 不/d 稀饭/n 黑色/n 金色/n 16/m g/eng 喜欢/v 摸索/v 金色/n 神机/n 过段时间/n 入手/v 小苹果/nr 金色/n 确实/ad 细致/a 送货/v 金色/n 漂亮/a 好看/v 不错/a 外面/f 便宜/a 一点/m 金色/n 好看/v 金色/n 小苹果/nr 不用说/l 超赞/v 查过/v 正宗/nz 国行/n 放心/v 金色/n 太/d 好看/v 不错/a 喜欢/v 金色/n 外壳/n 太潮/nr 金色/n 不是太好/l 买/v 送/v 老爸/n 喜欢/v 金色/n 不俗/a 买/v 金色/n 后悔/v 黑色/n 没用/v 小时/n 盖/v 花/v 金色/n 弱/a 金色/n 。。。
...	...

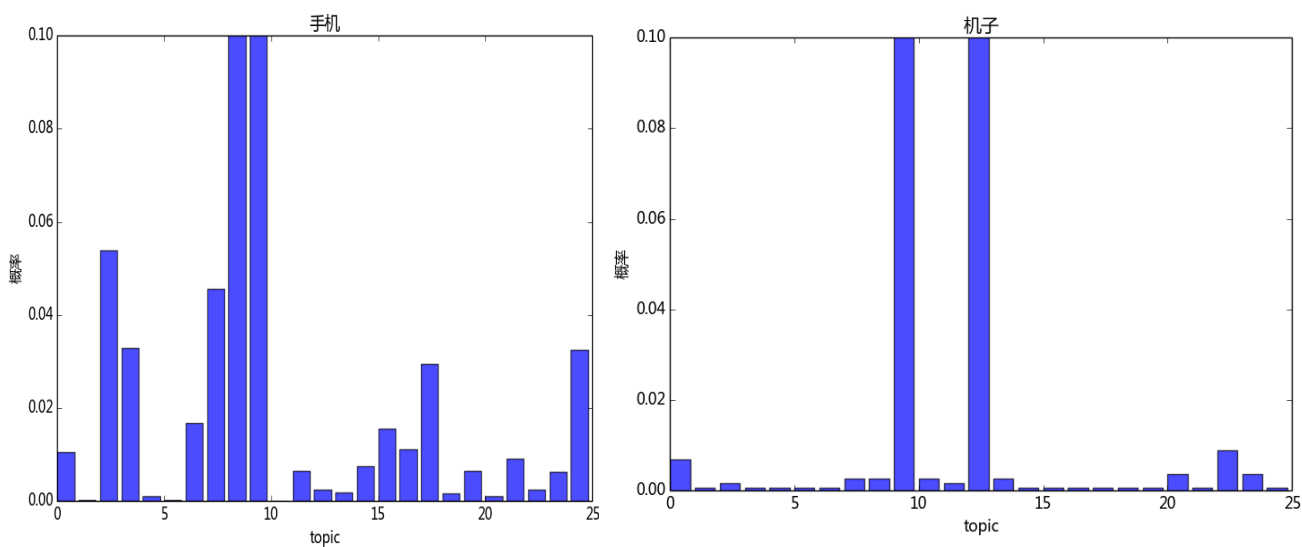
- GibbsLDA 主题抽取

在评价语言模型时候，通常使用困惑度。因而用 lda 抽取文档的 TOPIC 时，我们选择用困惑度来确定 TOPIC 数量，在数据集上，困惑度曲线如图所示。困惑度最小值位于 T=25 处，因此，参数 T 取 25。

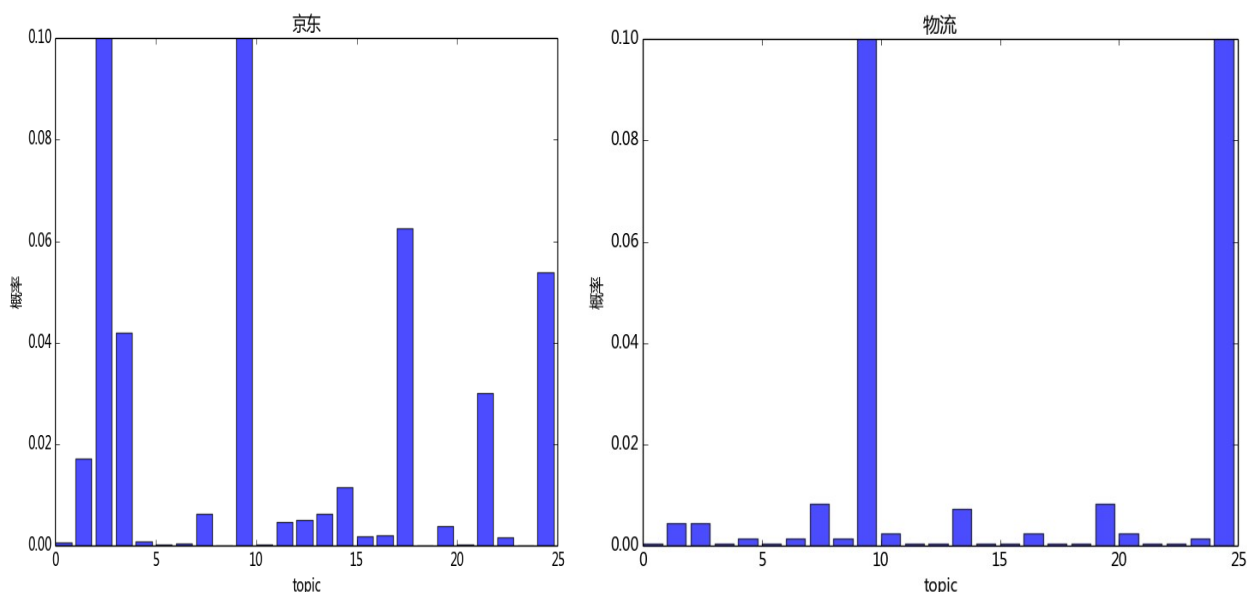


利用 LDA 抽取出的语言模型，对每一个名词进行推断，得到各名词的语义分布。

手机 VS 机子



京东 VS 物流



• 名词与名词间的相似度

通过 Gibbs 抽样，我们得到了共 170 个名词的语料集合的 TOPIC 分布。对每一个名词进行 TOPIC 推断，我们得到了 170 个名词以及该名词的语义分布。要利用名词的语义分布计算名词之间的相似程度，即计算离散分布之间的相似度，有两种较好的方法。一是余弦相似度[Steiyvers M, Griffiths T. Probabilistic topic models[J]. Handbook of latent semantic analysis, 2007, 427(7): 424-440.]，二是 KL-距离[Rosen-Zvi M, Griffiths T, Steiyvers M, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, 2004: 487-494.]。

A,余弦相似度。将每个点的概率值作为权重。

$$Similarity_{cosine} = \frac{\sum (X_i * Y_i)}{\sqrt{(\sum (X_i^2) + \sum (Y_i^2))}}$$

B.KL 距离。分别将每个名词的 TOPIC 分布视为基础模型 Q，将其他名词的 TOPIC 分布 P 与其比较。

$$Similarity_{kl}(N1, N2) = -\sum (P(x) * \log Q(x) + Q(x) * \log P(x))$$

在实验中，选取其中一种相似度。用标记数据测试相似度准确率，按正确率和召回率来选择哪一种相似度。（待实验。）

3，生成概念树。

1，对于初始树上的每一个结点 G (w_1, w_2, \dots, w_k)

$$p(w_k) = \frac{TF-IDF_k}{\sum (TF-IDF_i)} \quad \# \text{把每一个结点当作一篇文档，结点上的词按 TF-IDF 概率化。}$$

2，for G_i in InitialTree:

for n in noun:

$$p(n|w_k) = \frac{similarity(n, w_k)}{\sum (similarity(n_i, w_k))}$$

$$P(n|G)=p(n|w_1,w_2..w_k)=\sum p(n|w_k)*p(w_k)$$

3,更新树，结点为 $G=(l_1,l_2,l_3,...,l_k)=(p(l_1),p(l_2),p(l_3)...p(l_k))$

4,评论语料的所有词 $word=(w_1,w_2,w_3...w_m)$

名词的附属文档 $content_{Ni}=(p_1,p_2,p_3...p_n)$

以词频作为概率 $content_{Ni}=(p_{i1},p_{i2},p_{i3}...p_{in})=(TF_{i1},TF_{i2},TF_{i3}...TF_{in})$

for w in word:

$$p(w|G)=p(w|l_1,l_2,l_3,...,l_k)=\sum p(w|l_k)*p(l_k)$$

5,更新树，生成的树，每一个结点包含评论语料中所有的词以及词的概率。

步骤 3 更新树如下表（部分结点）：

结点	词-概率
根结点	
第一层结点	
第二层结点 1	
第二层结点 2	
第三层结点 3	

最终生成的概念树如下表（部分结点）：

结点	词-概率
根结点	
第一层结点	
第二层结点 1	
第二层结点 2	
第三层结点 3	

四，利用概念树对产品评论进行推断。