淘宝标题数据分析

印如意
yinruyi.hm@gmail.com

1 / 工作介绍

2 / 数据介绍

3 / 分析过程&实验

【贰】工作介绍

从产品名称中识别出主要的品牌和特征

按照品牌新（潮）、旧（2种），和K种常见款式，将所有产品都映射到2*K组0-1变量

从产品名称中计算名称的“信息量”

【贰】数据介绍

总体介绍

- maleTshirts/femaleTshirts
- 2800条/3413条
- 淘宝交易数据
- 07-15年

【贰】数据介绍

字段介绍

Id	11
Bouht_id	44074804657540
Item_id	44074804697540
Item_name	新款原单jumpingbeans纯棉咖底紫色花朵木耳边T恤
Dealtime	2010-08-09 00:01:00
original_price	23
special_price	23
item_url	http://item.taobao.com/item.htm?id=5530924286&_u=s4gqoibf631
Gender	1
Ageyear	0
Zipcode	110105
Shopname	leilei_5065
shop_url	http://store.taobao.com/shop/view_shop.htm?user_number_id=47820540
_merge	matched (3)

【贰】数据介绍

Item_name截取

dataset

蘑菇街2014新款韩版夏装女装复古青花瓷图案短袖女雪纺衫T恤女9.9
韩版休闲V领英文字母纯色宽松女士装款百搭短袖T恤大码打底衫
新款街头2013夏季白色五角星网纱背心2013年无袖女装黑色T恤圆领
清新气质Lizclaiborne基本款女装纯棉九分袖T恤有超加大码多色
羅蘭愛思LauraAshley原單绣花圆领修身T恤
小蚊子2014春款彩钻修身嘴唇T韩版百搭圆领短袖女款T恤#DX5260
【聚】不支持货到付款新宽松韩版休闲女士套装(印花T恤+休闲短裤)
i.t专柜潮牌文艺休闲百搭宽松vivi杂志原宿纯色短袖T恤女式t恤
〓胖乐园〓欧美休闲烧毁棉帅气简洁修身T恤两色大码
德国正品原单女式圆领蝙蝠短袖印花紫色宽松款上衣T恤有加大码
…

【叁】分析过程&实验从产品名称中识别出主要的品牌和特征

淘宝宝贝命名规则

- 名字写满30字，都由关键字组成。关键字之间不用空格和其他字符连接
- 真实相关，兼顾热搜关键词、成交关键词
- 去掉和宝贝无关的词
- 压缩优化，如“男士棉服”和“棉服男款”压缩成“男士棉服男款”
- 关键词优先级排序，重要放前面

【叁】分析过程&实验从产品名称中识别出主要的品牌和特征

识别品牌名称

- 粗分词（不加用户词典）

- 考虑：
- 品牌名=字典里的品牌名+错分的品牌名+英文品牌名

【叁】分析过程&实验从产品名称中识别出主要的品牌和特征

结果展示

- maleTshirt粗分词结果
- maleTshirt复合词结果
- maleTshirt标注品牌名/名词

- femaleTshirt粗分词结果
- femaleTshirt复合词结果
- femaleTshirt标注品牌名/名词

- allTshirt标注品牌名/名词

【叁】分析过程&实验从产品名称中识别出主要的品牌和特征

特征提取

- 细分词（加用户词典）

- 淘宝标题特性：
- 不是语句更像一堆形容词以及名词堆砌
- TF-IDF/筛选名词副词形容词
- n/nr/ns/nf/nt/nz/nl/ng/a/ad/an/ag/al

【叁】分析过程&实验从产品名称中识别出主要的品牌和特征

结果展示

- male特征未筛选词性
- male特征筛选名词形容词副词

- female特征未筛选词性
- female特征筛选名词形容词副词

- all特征未筛选词性
- all特征筛选名词形容词副词

【叁】分析过程&实验新旧品牌映射

问题

按照品牌新（潮）、旧（2种），和K种常见款式，将所有产品都映射到2*K组0-1变量

【叁】分析过程&实验新旧品牌映射

方法&结果

- male品牌-时间映射
- male品牌-特征映射

- female品牌-时间映射
- female品牌-特征映射

- all品牌-时间映射
- all品牌-特征映射

【叁】分析过程&实验新旧品牌映射

问题

- 新旧品牌难以区分

- 定义？新品牌/潮牌
- 数据量少

【叁】分析过程&实验信息量计算

问题

从产品名称中计算名称的“信息量”

【叁】分析过程&实验

方法

- 直接计算文本长度

- 信息熵 $\sum_{i=1}^{n} -p_i log p_i$

【叁】分析过程&实验

信息熵结果

- male全部结果 / male每个词信息量
- 结果

- 美国正品AF男士夏装短袖T恤A&F男圆领短袖T恤af短袖T恤afT恤夏装,1.5868116836
- 45.5元T恤秒杀T恤情侣装韩版修身圆领男装T恤/男士T恤/短袖T恤,1.42815496695

- 李宁雪铁龙赞助中国羽毛球比赛短袖T恤衫全国包申通可团购,0.169057562285
- 英国玛莎棉制花色亮片装饰短袖T恤衫,0.181645774141

- female全部结果 / female每个词信息量
- all全部结果 / all每个词信息量

Questions & Answers

THANKS

印如意

yinruyi.hm@gmail.com