基于行为数据
商品聚类

赵海臣

聚美优品
JUMEI.COM

背景

现有的商品聚类往往是基于商品的客观属性，例如价格、类别、品牌的精细划分，具有一定的客观合理性，但是没有考虑顾客的主观因素。作为电商企业，以顾客的主观因素进行商品类别划分更符合销售规律。

- 顾客的主观因素体现在顾客的行为中，例如浏览、购买、加购物车、收藏等等；
- 通过分析顾客的行为，通过分析顾客的行为大数据，对兴趣点进行聚类归纳，是一种更符合用户思维习惯的商品分类。
- 理论假设：
  - 每个用户都是有各自偏好与兴趣点的，若一个用户偏好于面膜，其浏览、购物、加购、收藏会有明显的面膜倾向；通过对大量用户进行分析，面膜相关的商品便将因为共同面膜兴趣的用户行为而能够得以聚类。
  - 用户的兴趣点是思维空间的一个吸引点，该吸引点能够泛化附近的相关商品，用户在浏览的时候，实际上是根据思维空间的吸引点以及对周围的搜索来行动的。因此，通过对大量用户的行为分析，我们可以将用户的思维解析出来，形成商品的基于人类思维的聚类。

基本技术路线

- 基础算法： k-means聚类衍生版
- 数据类型： 完全基于用户的行为记录
- 簇标签： 基于簇内商品标题分词词频，取最频繁特征词

聚美优品
JUMEI.COM

数据类型

- 商品聚类：
  - 理论依据：用户的购物偏好不一样，一个用户的浏览/购物是有品牌或者商品类别倾向的。因此，若两个商品被同一个用户浏览，则这两个商品间有关系的可能性较高。
    - 将用户的浏览记录按天的粒度进一步细分将提升基于用户行为的商品距离精确度，因为用户在一天中的浏览关注点往往更集中。
  - 基于用户近30天商品浏览记录ItemBrowseData进行分析
    - 清洗获得| userId | productId | rating |的分数表
  - 商品对应的维度特征：
    - 超高维稀疏用户维度：
      - 100,000,000+用户维度，以及对应的rating值
  - 数据处理方式：
    - 基于DataFrame稀疏矩阵存储与处理

基于行为数据的k-means

- 商品-商品距离公式选择：
  - 布尔型余弦相似度：

$$w_{uv} = \cos A = \frac{\vec{u} \cdot \vec{v}}{\| \vec{u} \| \cdot \| \vec{v} \|} = \frac{| N(u) \cap N(v) |}{\sqrt{| N(u) | | N(v) |}}$$

    - 由于一个商品对应的用户数量巨大，且用户的浏览次数大部分为区分不大的<10小值，用余弦相似度则只考虑用户是否浏览过的Boolean值，能够高效地利用大量的数据，并且运算量相对pearson等向量算法小。

基于行为数据的k-means

- 商品-商品距离公式选择：
  - 余弦相似度：

$$w_{uv} = \cos A = \frac{\vec{u} \cdot \vec{v}}{\| \vec{u} \| \cdot \| \vec{v} \|} = \frac{\sum Rating_{ui} * Rating_{vi}}{\sqrt{\sum Rating_{ui}^2 * \sum Rating_{vi}^2}}$$

基于行为数据的k-means

- 商品-商品距离公式选择：
  - Pearson相似度：

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E(X)^2}\sqrt{E(Y^2)-E(Y)^2}} = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{N}}\sqrt{\sum y^2 - \frac{(\sum y)^2}{N}}}$$

基于行为数据的k-means

- 初始质点的选择：
  - 从商品集中随机抽N个初始商品，并从行为数据中将这些商品相关的用户-商品交互数据提取出来，作为初始质点。
- 聚类迭代：
  - 计算各个商品与质点商品之间的距离
    - 余弦相似度
  - 簇更新
    - 每个商品取最近的质点作为簇标，相同簇标的商品为同一个簇
  - 质点更新(Key)
    - 簇各个维度均值
      - 簇中的每个维度(userId)值加起来(布尔值True = 1 / False = 0)，再除以簇中点的个数
    - 簇非零维度个数
      - 簇中商品的平均浏览用户数averageUserAmount
    - 簇更新策略
      - 取簇中维度均值最高的averageUserAmount个用户维度作为新的簇质心的坐标!=0/True，其它维度皆置0/False

迭代过程

- 空间搜索过程
  - 基于行为的商品聚类是在一个极其广袤的超高维稀疏空间中的聚类过程，此例中一个商品对应的用户维度>100,000,000。
  - 余弦相似度的一个核心点在于商品-商品之间有共同的非0中介维度，否则二者距离为0；因此，若一个商品与所有其它商品都没有任何交集，则将被自动删减。
  - 簇更新策略中限制每个簇心的有效非0维度数为簇中商品平均浏览用户数，能够在维度空间进行自动的搜索，防止维度发散，以解决超高维空间的稀疏问题。
  - 由于簇心的非0维度个数取的簇平均值，簇心将首先在维度空间搜索商品聚集的维度，再基于距离计算，在对应的聚集维度进行聚类。

评估指标

- 基础评估指标：
  - 平均距离：
    - sigma(1-被聚类的点与簇心的余弦相似度)/所有被聚类点的个数
  - 收敛速度：
    - 上次平均距离 - 此次平均距离 / 上次平均距离
  - 点覆盖率：
    - 成功聚类的点个数 / 样本集中所有的点个数

降噪与计算优化

- 只考虑高交互商品
  - 基于用户行为的聚类完全依赖于用户-商品间的交互，若一个商品与用户交互记录较少，则对该低交互商品基于行为聚类是不可靠的，根据余弦相似度，低交互商品反而容易获得较高的近似分，因而容易引入强干扰。
    - 对商品的交互次数进行排序，只取前1000000个高交互商品进行聚类。
      - 能够大幅提升聚类结果准确性
      - 能够提升运算速度
- 抽样进行聚类
  - 随机抽样n个商品作聚类样本商品，再从行为数据中将该n个商品的交互记录取出
    - 抽样可以有效地提高计算速度，并且可以规避很多噪声数据。若抽样率10%，则单噪点被过滤的概率为90%。
    - 完全随机抽样可以保存大部分密集簇，因此如果抽样点个数仍远大于目标簇个数，簇的聚类准确性不会有较大的影响。
  - 基于抽样点聚类完获得簇心后，假设簇心是完美的，则使用簇心对所有商品重新进行聚类簇划分，即完成对所有点的聚类。

改进的渐进式聚类

考虑到若初始簇心过于密集，将导致簇与簇之间的相互干扰，比如两个簇心比较靠近，将导致簇的分裂。

改进方案：

- 将总数为n的目标簇个数，等分为m次进行聚类，每次迭代循环完成n/m个簇的聚类。
- 每次聚类完成之后，假设距离簇心<l的是该簇的点，则从训练集中删减掉距离簇心距离<l的点，防止对后续聚类的干扰。
- 如此每次聚类完成后，密集的簇以及对应的点被移除，能强制后续的簇心搜索稀疏的、未聚类的区域，防止簇心距离过近争夺分裂同一个簇。
- 每轮聚类完成，将获得的簇心保存下来，最后合成整体簇心集合。
- 在通过渐进聚类获得每一轮的簇心后，由于每一轮关注的都是局部的，因此最终需要将所有簇心集合一起进行全局迭代聚类，以纠正局部偏差。

miniBatch

- miniBatch可以通过抽样技术，加快聚类速度
  - 每一轮迭代中，都从样本集中随机抽取一个更小样本集，通过在小样本集上进行簇划分与簇心更新，大幅提高运算速度。

商品簇自动生成标签

商品名称中包含了这个商品的最浓缩的信息，因此通过对商品名称进行分词，获得每个商品的特征，再通过统计每个簇中商品的词频，每个簇选出词频最高的10个词作为簇标签。

- 分词精准性提高策略以及噪声过滤
  - 将商品库中的商品名称、分类名称添加到分词工具自定义字典中，防止误分与漏分
  - 对于分词后的汉字词组，字数小于等于2的词过滤，因品牌与功效大部分大于2个汉字
  - 对于分词后英文单词+数字，长度小于等于4的词过滤，因英文品牌大部分大于4个字母或数字
  - 过滤掉非汉字、英文、数字的分词结果(标点)
- 只基于词频TermFrequency，而不加上InversedDocFreqency，原因是商品标题中基本上都是浓缩后的有效精华词，IDF反而将破坏一些普遍功效性描述关键词，例如洗发水、美白之类的。

结论

- 基于用户行为的商品聚类发现：
  - 品牌类簇：知名品牌的商品往往是基于品牌的簇，即簇中的商品往往是同一个品牌的；
  - 类别簇：例如指甲钳、咖啡、绿茶之类的商品，往往是类别簇，即簇中的商品往往是相同类别的；
  - 噪声异常簇：即未成功聚类的簇，如同簇与簇的边界，混杂着各种有明显区别商品类别，或者商品与商品间并没有明显的共同特征。
    - 噪声异常簇可以通过进一步的处理，例如进一步划分，或者对异常簇簇心进行删减来改善。
    - 根据聚类过程可知，异常簇中的商品仍然是相互之间有一定联系的商品。

类别簇与簇标例子

centroid	word	count	cluster_size	row_number
3143941	美国	0.8764160659114315	971	1
3143941	女士	0.7775489186405767	971	2
3143941	斜挎包	0.388259526261586	971	3
3143941	瑞贝卡	0.29866117404737386	971	4
3143941	明可弗	0.29042224510813597	971	5
3143941	单肩	0.2646755921730175	971	6
3143941	时尚	0.23171987641606592	971	7
3143941	rebecca	0.22245108135942326	971	8
3143941	凯特	0.19979402677651906	971	9
3143941	丝蓓	0.19979402677651906	971	10

productId	centroid	distance	product_short_name
3289366	3143941	0.17081486713073674	美国•rebecca minkoff 瑞贝卡明可弗 女士小号斜挎包 1169926877
3423588	3143941	0.16093484986175585	美国•汤丽柏琦(Tory Burch) 时尚黑色皮质女士单肩斜挎包18169277-001
3730818	3143941	0.16049907891132242	美国•汤丽柏琦(Tory Burch) 时尚棕色皮质女士单肩斜挎包39053-209
3602720	3143941	0.15931737313308109	美国•汤丽柏琦(Tory Burch) 时尚黑色皮质女士单肩斜挎链条包37594-001
3514416	3143941	0.1569316614538851	美国•rebecca minkoff瑞贝卡.明可弗 女士斜挎包 1176393847
3730836	3143941	0.15331103516796663	美国•汤丽柏琦(Tory Burch) 时尚棕色皮质女士斜挎单肩包11169702-211
3512444	3143941	0.1530710701420025	美国•rebecca minkoff瑞贝卡.明可弗 女士流苏斜挎包 1176393893
3004462	3143941	0.15114173098063566	美国•REBECCA MINKOFF/瑞贝卡.明可弗 时尚简约链条单肩女包
3602686	3143941	0.15039458033972927	美国•汤丽柏琦(Tory Burch) 时尚黑色皮质女士单肩斜挎链条包34029-001
3730816	3143941	0.1494673588682159	美国•汤丽柏琦(Tory Burch) 时尚黑色皮质女士单肩斜挎包33838-001
3730522	3143941	0.1492220039998552	美国•汤丽柏琦(Tory Burch) 时尚黑色皮质女士单肩斜挎包32196-001
3730850	3143941	0.14816194208498976	美国•汤丽柏琦(Tory Burch) 时尚黑色皮质女士单肩斜挎包35620-001
2534612	3143941	0.14782809899727065	美国•REBECCA MINKOFF 瑞贝卡.明可弗 单肩斜挎女包
3692782	3143941	0.14782809899727065	美国•Rebecca 蓝色皮质女士单肩斜挎包HF35ELVX08
3526468	3143941	0.14634235175124088	美国•rebecca minkoff瑞贝卡.明可弗 女士迷你斜挎包 1183973069
3289386	3143941	0.14598637372036238	美国•rebecca minkoff 瑞贝卡.明可弗 瑞贝卡明可弗 女士小号斜挎包 1169926880
3179844	3143941	0.139159082112814	美国•REBECCA MINKOFF/瑞贝卡.明可弗 绗缝纹旋转钮扣女士单肩斜挎包
3687704	3143941	0.1379728923974526	美国•rebecca minkoff瑞贝卡.明可弗2017官方同款新款 LOVE系列V纹衍缝小号斜挎包 HSP7GLVX45-A
2230138	3143941	0.1374773395873622	美国•Rebecca Minkoff女士海蓝色纯皮菱格单肩包斜挎包H324I001
3219879	3143941	0.13655774839978377	美国•汤丽柏琦(Tory Burch) 时尚黑色皮质女士单肩斜挎链条包34037-001
3291040	3143941	0.13655774839978377	美国•Rebecca Minkoff 瑞贝卡.明可弗 女士拉链小号斜挎包 1171381651
3513750	3143941	0.13655774839978377	美国•rebecca minkoff 瑞贝卡.明可弗 女士黑色铆钉M.A.C系列链条包 1045044845
3514244	3143941	0.13608679574904795	美国•rebecca minkoff瑞贝卡.明可弗 女士斜挎包 1176393842
2230124	3143941	0.13588932662615225	美国•Rebecca Minkoff 石头色纯皮时尚铁链斜挎包HF35MFCX01
3179862	3143941	0.13562687355178338	美国•REBECCA MINKOFF/瑞贝卡.明可弗 菱格纹翻盖旋锁链条女士斜挎包
3516144	3143941	0.13517414889732077	美国•rebecca minkoff 瑞贝卡.明可弗瑞贝卡明可弗 女士流苏小号斜挎包 1161363407
2699235	3143941	0.13291555207098021	美国•Rebecca Minkoff 白色纯皮女士UNLINED FEED BAG系列单肩斜挎包HS16IULX62
3179834	3143941	0.1314929728380438	美国•REBECCA MINKOFF/瑞贝卡.明可弗 流苏翻盖 女士斜挎包
2737706	3143941	0.13077638454490423	美国•Tory Burch(汤丽柏琦) 黑色纯皮女士单肩手提包 31159052-001
3526382	3143941	0.12955055512625914	美国•rebecca minkoff瑞贝卡.明可弗 女士斜挎包 1183659741
3526460	3143941	0.12955055512625914	美国•rebecca minkoff瑞贝卡.明可弗 女士斜挎包 1183660000
3531952	3143941	0.1290728893976494	美国•rebecca minkoff瑞贝卡.明可弗 女士迷你斜挎包 1183659793
3730822	3143941	0.12766447395858072	美国•汤丽柏琦(Tory Burch) 时尚黑色皮质女士单肩斜挎包36731-001
3526416	3143941	0.12752571097814502	美国•rebecca minkoff瑞贝卡.明可弗 女士斜挎包 1185100977
3219877	3143941	0.12635334150137517	美国•汤丽柏琦(Tory Burch) 时尚酒红色皮质女士单肩斜挎链条包31409-517
2922144	3143941	0.12403473458920847	美国•REBECCA MINKOFF 瑞贝卡.明可弗 女士流苏翻盖单肩斜挎包
2935559	3143941	0.12340670658118719	美国•Rebecca Minkoff 玫红色十字纹皮女士单肩斜挎包HP36ESSX10
3658304	3143941	0.12329715777434853	美国•凯特·丝蓓 Kate Spade女款 avva系列斜跨包单肩包 黑色
3153868	3143941	0.12209073054533606	美国•Rebecca Minkoff/瑞贝卡.明可弗 时尚流苏铆钉装饰斜挎包
3526364	3143941	0.12205795890857729	美国•rebecca minkoff瑞贝卡.明可弗 女士斜挎包 1183659846
3476884	3143941	0.12190217332000222	美国•rebecca minkoff瑞贝卡.明可弗 女士斜挎包 1176393909
3598492	3143941	0.1210166964639644	美国•汤丽柏琦(Tory Burch) 时尚绿色皮质女士单肩斜挎链条包37594-307
3486524	3143941	0.1207011373963169	美国•ALEXANDER WANG 亚历山大王 女款黑色牛皮手提单肩包 20R0345 001
3512464	3143941	0.11976481509408357	美国•rebecca minkoff瑞贝卡.明可弗 女士迷你挎包 1176393917
3514032	3143941	0.1194880298498108	美国•rebecca minkoff瑞贝卡.明可弗 女士斜挎包 1178895378
3526392	3143941	0.11929897998982567	美国•rebecca minkoff瑞贝卡.明可弗 女士斜挎包 1183659934
2922137	3143941	0.11861392864779444	美国•REBECCA MINKOFF 瑞贝卡.明可弗 女士流苏单肩斜挎马鞍包
3289492	3143941	0.11826247919781652	美国•rebecca minkoff瑞贝卡.明可弗 女士小号钉珠斜挎包 1171381604
3560034	3143941	0.11826247919781652	美国•rebecca minkoff 瑞贝卡.明可弗 女士杏色流苏迷你斜挎包 1178895373
3730820	3143941	0.11826247919781652	美国•汤丽柏琦(Tory Burch) 时尚蓝色皮质女士单肩斜挎包39091-433
3560294	3143941	0.11800054647490729	美国•rebecca minkoff 瑞贝卡.明可弗 女士白色斜挎包链条包 1156452360
3467872	3143941	0.11800054647490729	美国•botkier柏柯尔 女士酒红色斜挎包 1159769698
3717058	3143941	0.11759139445536937	美国•rebecca minkoff 瑞贝卡.明可弗 女士黑色手拿斜挎两用包 HH16EDSC45-BLACK001
2737618	3143941	0.1169699542776235	美国•Tory Burch(汤丽柏琦) 黑色纯皮女士单肩斜挎包12169054-001
3630674	3143941	0.11596590176067832	美国•rebecca minkoff瑞贝卡.明可弗 女士黑色斜挎包 1178895269
3179849	3143941	0.11498327637597498	美国•REBECCA MINKOFF/瑞贝卡.明可弗 长条流苏拉链方形女士手提斜挎包

品牌簇与簇标例子

centroid	word	count	cluster_size	row_number
1889879	美国	0.92	100	1
1889879	伊丽莎白雅顿	0.91	100	2
1889879	银级	0.27	100	3
1889879	保湿	0.26	100	4
1889879	金致	0.18	100	5
1889879	绿茶	0.17	100	6
1889879	精华液	0.13	100	7
1889879	胶囊	0.13	100	8
1889879	调理	0.13	100	9
1889879	眼霜	0.12	100	10

productId	centroid	distance	product_short_name
588703	1889879	0.25361765957534205	美国•伊丽莎白雅顿 (Elizabeth Arden)金致胶囊精华液 14ml
2520043	1889879	0.243834942540459	美国•伊丽莎白雅顿(Elizabeth Arden)超时空金致礼盒装（金致（导航）眼部胶囊精华液10.5ml+金致（导航）眼部胶囊精华液1.2ml*2）
1035	1889879	0.23328667948016923	美国•伊丽莎白雅顿 (Elizabeth Arden)金致导航眼部胶囊精华液 10.5ml（60粒）
1576	1889879	0.22783326630922424	美国•伊丽莎白雅顿 (Elizabeth Arden)金致胶囊精华液28ml（60粒）
2462580	1889879	0.21613871576659996	美国•伊丽莎白雅顿 (Elizabeth Arden)新生代眼部精华液 15.6ml（90粒）
312824	1889879	0.209749678823856	美国•伊丽莎白雅顿 (Elizabeth Arden)护肤套装（保湿调理露 200ml+复合霜 75ml）
328	1889879	0.20765994249237243	美国•伊丽莎白雅顿 (Elizabeth Arden)复合面霜（显效复合霜） 75ml
2454444	1889879	0.20708639728118064	美国•伊丽莎白雅顿 (Elizabeth Arden)24小时水感恒润套装（眼霜15ml+保湿霜50ml+保湿乳50ml）
2600153	1889879	0.20437086825503897	美国•伊丽莎白雅顿(Elizabeth Arden)静享岁月礼盒（保湿调理露200ml+复合（面）霜75ml/70g+金致导航眼部胶囊精华液1.2ml+无瑕未来活颜晚霜
1270960	1889879	0.20007087978212915	美国•伊丽莎白雅顿 (Elizabeth Arden)银级（爽肤水）柔肤水 150ml
1650	1889879	0.19604044973530454	美国•伊丽莎白雅顿 (Elizabeth Arden)（柔润）保湿调理露 200ml
1271028	1889879	0.19541155811321773	美国•伊丽莎白雅顿 (Elizabeth Arden) 金致胶囊精华液 3.2ml
539298	1889879	0.19138278668491848	美国•伊丽莎白雅顿 (Elizabeth Arden) 基础护理三件套（双效洁肤露150ml+保湿调理露200ml+显效复合霜75ml）
303	1889879	0.1913552081043596	美国•伊丽莎白雅顿 (Elizabeth Arden)新生代胶囊精华液 41.4ml （90粒）
2714	1889879	0.186181359156735	美国•伊丽莎白雅顿 (Elizabeth Arden)银级日霜（日间面霜） 75ml
2520042	1889879	0.18573548269311535	美国•伊丽莎白雅顿(Elizabeth Arden)金致闪耀礼盒装（金致胶囊精华液14ml+金致胶囊精华液7粒装3.2ml*2）
2710	1889879	0.18440126096695128	美国•伊丽莎白雅顿 (Elizabeth Arden)银级（护肤）三件套（银级日霜75ml+银级晚霜50ml+银级眼霜15ml）
452	1889879	0.17445547998153899	美国•伊丽莎白雅顿 (Elizabeth Arden)水感恒润持久保湿乳50ml(24小时持久保湿乳50ml)
1804645	1889879	0.1707828593845622	美国•伊丽莎白雅顿 (Elizabeth Arden)银级调理四件套（银级日霜75ml+银级晚霜50ml+银级眼霜15ml+调理露200ml）
1804658	1889879	0.17071399995320932	美国•伊丽莎白雅顿 (Elizabeth Arden)银级滋养两件套（银级眼霜15ml+滋养霜50ml）
1867068	1889879	0.17017441936977434	美国•伊丽莎白雅顿 (Elizabeth Arden)银级护肤四件套（银级日霜75ml+银级晚霜50ml+银级眼霜15ml+爽肤水150ml）
2927469	1889879	0.17015559356915413	美国•伊丽莎白雅顿 (Elizabeth Arden)复合面霜 75ml*2
373	1889879	0.16774731998384743	美国•伊丽莎白雅顿 (Elizabeth Arden)洁净爽肤水 200ml
313	1889879	0.16470812699884985	美国•伊丽莎白雅顿 (Elizabeth Arden)晚安好眠滋养(乳)霜 50g
169	1889879	0.15960423512421093	美国•伊丽莎白雅顿 (Elizabeth Arden)水感恒润持久保湿霜 50ml
2005598	1889879	0.15561401394507818	美国•伊丽莎白雅顿 (Elizabeth Arden)金致（导航）眼部胶囊精华液 1.2ml
1804638	1889879	0.1545690111751248	美国•伊丽莎白雅顿 (Elizabeth Arden)银级日晚调理三件套（银级日霜75m+晚间面霜 50ml+调理露200ml）
1665712	1889879	0.1472788348651742	美国•伊丽莎白雅顿 (Elizabeth Arden)显效护理三件套（洁面乳150ml+爽肤水200ml+复合霜75ml）
153	1889879	0.14714271734398446	美国•伊丽莎白雅顿 (Elizabeth Arden)银级（护理）眼霜15ml
2229	1889879	0.14615543036390813	美国•伊丽莎白雅顿 (Elizabeth Arden)微粒保湿/保湿微粒洁面乳 150ml
182	1889879	0.14251269487284812	美国•伊丽莎白雅顿 (Elizabeth Arden)银级晚霜/晚间面霜 50ml
1867087	1889879	0.14044838999751558	美国•伊丽莎白雅顿 (Elizabeth Arden)银级滋润两件套（银级眼霜15ml+爽肤水150ml）
589	1889879	0.13911599840957634	美国•伊丽莎白雅顿 (Elizabeth Arden)水感恒润持久保湿眼霜（24小时持久保湿眼霜） 15ml
1867071	1889879	0.138899900038948	美国•伊丽莎白雅顿 (Elizabeth Arden)银级日晚三件套（银级日霜75ml+银级晚霜50ml+爽肤水150ml）
1465546	1889879	0.13437437719170486	美国•伊丽莎白雅顿(Elizabeth Arden)银级日晚护肤两件套（银级日霜75m+晚间面霜 50ml）
657813	1889879	0.11949161589026079	美国•伊丽莎白雅顿 (Elizabeth Arden)金致乳液SPF30 PA++ 50ml
374	1889879	0.11895658175235789	美国•伊丽莎白雅顿 (Elizabeth Arden)双效洁肤露 150ml(绿色)
2683185	1889879	0.11761603100489774	美国•伊丽莎白雅顿(Elizabeth Arden)经典轻感平衡油 100ml
1703022	1889879	0.11736484962941093	美国•伊丽莎白雅顿 (Elizabeth Arden)保湿护眼两件套（洁面乳150ml+保湿眼霜15ml）
3061675	1889879	0.10801270204083814	美国•伊丽莎白雅顿 (Elizabeth Arden) 润肤调理套装（保湿调理露 200ml+复合面霜 75ml+金致胶囊精华液3.2ml+时空纯净清洁霜30ml+拎袋）
294	1889879	0.10034302379078759	美国•伊丽莎白雅顿 (Elizabeth Arden)净柔/纯净眼唇卸妆液 100ml
2600267	1889879	0.09623157120733734	美国•伊丽莎白雅顿(Elizabeth Arden)臻心温润礼盒(双效洁肤露150ml+保湿调理露200ml+水感保湿眼霜15ml)
28655	1889879	0.09166260368339466	美国•伊丽莎白雅顿 (Elizabeth Arden)经典润泽防护霜 50ml
387704	1889879	0.08944704463729958	美国•伊丽莎白雅顿 (Elizabeth Arden)绿茶蜜滴舒体霜/身体霜 250ml
9339	1889879	0.08891751485723569	美国•伊丽莎白雅顿 (Elizabeth Arden) 绿茶蜜滴舒/身体霜 400ml
1700955	1889879	0.08801577824250172	美国•伊丽莎白雅顿 (Elizabeth Arden)保湿调理三件套（洁面乳150ml+调理露200ml+植物面膜100ml）
2600295	1889879	0.08559931344748843	美国•伊丽莎白雅顿(Elizabeth Arden)滋润守护礼盒(水感保湿霜 50ml+水感保湿眼霜 15ml+水感保湿乳 50ml+时空纯净清洁霜30ml)
1270961	1889879	0.08473736808248511	美国•伊丽莎白雅顿 (Elizabeth Arden)经典身体润泽霜 200ml
1417535	1889879	0.07934422065511058	美国•伊丽莎白雅顿 (Elizabeth Arden)第五大道滋润身体乳 200ml
365	1889879	0.078871030524332	美国•伊丽莎白雅顿 (Elizabeth Arden)保湿植物面膜 100ml
715220	1889879	0.07865882857336776	美国•伊丽莎白雅顿 (Elizabeth Arden)晶璨皙颜菁华水 200ml
2187690	1889879	0.0772035441777504	美国•伊丽莎白雅顿 (Elizabeth Arden)绿茶香体套装（蜜滴舒体霜400ml+忍冬香氛50ml）
28680	1889879	0.07541914453042066	美国•伊丽莎白雅顿 (Elizabeth Arden)水颜清爽保湿乳 50ml
1889879	1889879	0.0731094168349909	美国•伊丽莎白雅顿 (Elizabeth Arden)晶璨皙颜面霜 50ml
146977	1889879	0.06881629424598937	美国•伊丽莎白雅顿 (Elizabeth Arden)绿茶沐浴啫喱/露 500ml

聚类Ensemble

- 将同一个商品在两个聚类结果(例如使用不同的距离公式、不同的聚类方式)的质心名称连接起来，即完成Ensemble
  - 两个商品在两个聚类中都在同一个簇的商品，因两个聚类结果中簇名一样，因此将最终同样出现在同一个簇中；
  - 两个商品在一个聚类结果中在同一个簇，但在另一个聚类中在不同簇中，则将被分到不同簇中；
  - 若一个商品在两个聚类结果中都没有共同的伙伴，将被分到单独的一个簇标，该商品将被判断成噪点，而被抛弃。

- Ensemble的结果进一步大大提升了聚类的纠错效果，并且将不同条件下的聚类结果进一步以交集形式细分，例如单独聚类获得了“雅诗兰黛”的簇，在Ensemble后将进一步获得“雅诗兰黛面霜”、“雅诗兰黛粉底”等更细的类别。

不同距离公式Ensemble

- pearson距离公式主要考察向量相似度
- boolean距离公式主要考察item相似度
- 二者Ensemble得到的结果中，既有item相似又有向量相似，所以pearson boolean Ensemble能获得较为优秀的结果

- Ensemble弊端：很容易导致碎片化，虽然单簇很纯净，但是簇都很小，并且很容易导致单点簇。

THE END

THANK YOU!