作业三 挖掘用户关注点

【任务】

从用户评论中挖掘用户关注点，对于改进产品设计以及产品营销都具有积极意义。请同学们针对淘宝平台中部分商品的用户评论数据，分别使用文本关键词提取方法以及主题模型，提取其中的用户关注点。

（1）关键词提取：对于所有数据，进行文本关键词提取，挖掘该品类整体上用户所关注的焦点。目前，用于文本关键词提取的主要方法有四种：基于 TF-IDF 的关键词抽取、基于 TextRank 的关键词抽取、基于 Word2Vec 词聚类的关键词抽取以及多种算法相融合的关键词抽取。

（2）主题模型：面对浩如烟海的文档，把相似的文章聚合起来，并且提取描述聚合后主题的重要关键词，通过主题模型，挖掘在该品类中用户关注的几个主要话题及对应的话题内容。主题提取模型从 LSA 发展到了 PLSA，再到当前最常用的 LDA 以及在此基础上的 lda2vec 等改进模型。

同学们可根据数据集特点自行选取方法或进行创新改进。

【数据来源】

本作业附件中提供了食品、服装及化妆品三个品类的用户评论数据集，同学们可以任选其中一类完成即可。此外，如果认为样本数据不足，鼓励个人自行从淘宝平台爬取数据进行补充（食品、服装品类数据集中提供了商品链接，作业附件中提供了之前的淘宝评论数据爬取代码及流程介绍，可根据需求自行修改，仅供参考）。

【作业提交】

请提交以下文件:

一、作业报告（10 页以内，精要，写你认为重要的内容），内容须包括:

- 数据描述
- 数据处理流程简述（如有数据清洗、筛选等过程，可简单说明）,使用的模型方法/工具
- 结果分析

- 问题思考：本任务可应用的场景及意义
- 总结：不足与展望

二、所用数据及可运行源代码文件

（若进行了数据补充，请附上相应数据及代码，并在报告中进行说明）

打包发送至邮箱【njubgw@126.com】

所有文件及邮件命名方式【学号 姓名】

截止日期【8 月 20 日 24 点前】