# 大数据处理与分析课程项目设计报告

|  |  |  |  |
| --- | --- | --- | --- |
| 学校 | 北京石油化工学院 | 专业 | 数据科学与大数据技术 |
| 班级 | 大201/大202 | 教师 | 杨鲲鹏 |
| 考核方式 | 分组完成 | 时间 | 2022/6/30 |
| 姓名 | 杨发元 刘文龙  刘好 褚英皓 | 学号 | 2020310819 2020310812  2020310811 2020310863 |
| 项目分工 | 杨发元：撰写报告 刘文龙：调试代码 撰写报告 项目答辩  刘好：制作PPT 调试代码 褚英皓：项目答辩 | | |
| 项目标题 | 用户情感可视化分析 | | |
| 项目目标 | 以网络舆情分析为背景，要求根据用户的评论来对品牌的议题进行数据分析与可视化，并应当通过数据可视化图表和数据分析方法，对感兴趣的内容进行探索性数据分析。 | | |
| 设计思路 | 将下载好的数据通过WinSCP工具上传到虚拟机，在虚拟机上运行Hadoop，用Hadoop集群做数据处理，将处理好的数据再通过python做出可视化。使用条形图，热力图和饼状图，清楚地表明各种数量的多少，用于查看总体的情况、发现异常值、显示多个变量之间的差异，以及检测它们之间是否存在任何相关性。 | | |
| 实施过程 | **一、环境准备**  【步骤一】  在虚拟机上解压JDK并安装，然后做好JDK的环境配置。    图1  【步骤二】  安装Hadoop，做好Hadoop的环境配置，确定已安装成功后做Hadoop的集群部署。    图2  【步骤三】  将配置好的虚拟机进行克隆，并命名为hadoop01、hadoop02、hadoop03。使3个虚拟机之间的网络能够相互ping通后，进行格式化。格式化成功以后就可以启动所有进程。      图3  **二、数据处理**  【步骤一】  引入相关库 import numpy as np  import pandas as pd  from pylab import \*  import matplotlib.pyplot as plt  import seaborn as sns  import jieba  【步骤二】  读取数据，基础分析数据 *#读取数据*  df = pd.read\_csv('./data/earphone\_sentiment.csv')  *#查看数据前几行信息*  df.head()    图4  #查看数据基本信息  print(df.info())  print("----------------")  #查看数据缺指信息  print(df.isnull().sum())  print("----------------")  print(df['sentiment\_word'].unique())  print("----------------")  print(df['sentiment\_value'].value\_counts())  输出如下，主要分析了表格中缺省值和sentiment\_word这一列的取值，因为我们这里分析主要也是针对情感词来进行分析的：  <class 'pandas.core.frame.DataFrame'>  RangeIndex: 17176 entries, 0 to 17175  Data columns (total 5 columns):  # Column Non-Null Count Dtype  --- ------ -------------- -----  0 content\_id 17176 non-null int64  1 content 17176 non-null object  2 subject 17176 non-null object  3 sentiment\_word 4966 non-null object  4 sentiment\_value 17176 non-null int64  dtypes: int64(2), object(3)  memory usage: 671.1+ KB  None  ----------------  content\_id 0  content 0  subject 0  sentiment\_word 12210  sentiment\_value 0  dtype: int64  ----------------  dtype: int64  0 12210  1 4376  -1 590  Name: sentiment\_value, dtype: int64  ----------------  不评价 12210  好 3302  不错 569  差 415  强 244  牛 133  垃圾 32  高级 31  追求 26  呵呵 24  难听 22  骗 21  噪音 16  用心 16  舒适 16  疼 14  音染 14  水准 12  轰 10  精致 7  惊艳 7  良心 7  无语 6  无奈 5  不舒服 4  小巧 3  充足 3  上当 2  辣鸡 2  模糊 2  混浊 1  Name: sentiment\_word, dtype: int64  【步骤三】  空值处理，数据映射  由输出的数据可以分析出，我们可以根据sentiment\_value的值对sentiment\_word进行划分为3类，分别为0对于没有评价，我们这里当作是中性评价，1对于好评，-1对应差评。再做数据预处理，填上空值，对sentiment\_value做映射。  # 填上空值  df['sentiment\_word'].fillna('不评价', inplace=True)  # 将sentiment\_value映射  map\_sentiment\_value = {-1: '差评', 0: '中评', 1: '好评'}  df['sentiment\_value'] = df['sentiment\_value'].map(map\_sentiment\_value)  做个关于情感词和主题之间的透视表，可以很直观的看出，每个主题中好评、中评、差评的情况。  # 做sentiment\_value的透视表  df\_pivot\_table = df.pivot\_table(index='subject', columns='sentiment\_value', values='sentiment\_word',  aggfunc=np.count\_nonzero)    df\_r\_pivot\_tabel = df.pivot\_table(index='sentiment\_value', columns='subject', values='sentiment\_word',  aggfunc=np.count\_nonzero)  sentiment\_value 中评 好评 差评  subject  价格 495 256 42  其他 9493 2837 326  功能 83 63 10  外形 85 68 5  舒适 10 37 22  配置 1452 759 121  音质 592 356 64  subject 价格 其他 功能 外形 舒适 配置 音质  sentiment\_value  中评 495 9493 83 85 10 1452 592  好评 256 2837 63 68 37 759 356  差评 42 326 10 5 22 121 64  【步骤四】  对评论进行分词分析  对数据中content评论部分进行分析，而评论中有许多不必要的内容，需要对此分词并去除停用词（也就是一些连接词，以帮助分析)，对评论分词，主要应用jieba这个分词包。停用词用的是常用的中文停用词。  stopwords = []  with open('./data/mStopwords.txt', encoding='utf-8') as f:  for line in f:  stopwords.append(line.strip('\n').split()[0])    # 切词  rows, cols = df.shape  cutwords = []  for i in range(rows):  content = df['content'][i]  g\_cutword = jieba.cut\_for\_search(content)  cutword = [x for x in g\_cutword if (len(x) > 1) and x not in stopwords]  cutwords.append(cutword)    s1 = pd.Series(cutwords)  df['cutwords'] = s1  print(s1.value\_counts())  不同情感的分析，所以我们根据sentiment\_value的值，将表格划分为差评，中评和好评三个表。  # 根据情感划分切词  df\_pos = df.loc[df['sentiment\_value'] == '好评'].reset\_index(drop=True)  df\_neu = df.loc[df['sentiment\_value'] == '中评'].reset\_index(drop=True)  df\_neg = df.loc[df['sentiment\_value'] == '差评'].reset\_index(drop=True)  **三、处理结果**  【步骤一】  柱状图  柱状图主要分析不同主题下，不同情感评论的评论数，来看用户重点关注的地方是什么。  df\_pivot\_table.plot.bar()  for x, y in enumerate(df\_pivot\_table['中评'].values):  plt.text(x - 0.2, y, str(y), horizontalalignment='right')  for x, y in enumerate(df\_pivot\_table['好评'].values):  plt.text(x, y, str(y), horizontalalignment='center')  for x, y in enumerate(df\_pivot\_table['差评'].values):  plt.text(x + 0.2, y, str(y), horizontalalignment='left')  plt.title('不同主题，不同情感评论数')  plt.ylabel('评论数')  plt.show()  然后我们可以看看在褒义词和贬义词中比较多的评论在说什么。  # 取出褒义词和贬义词中的评论列  pos\_word = df\_pos['sentiment\_word'].value\_counts()  neg\_word = df\_neg['sentiment\_word'].value\_counts()    pos\_word.plot.barh()  for x, y in enumerate(pos\_word):  plt.text(y, x, str(y), horizontalalignment='left')  plt.xlabel('评论数')  plt.title('褒义词评论词统计')  plt.show()    neg\_word.plot.barh()  for x, y in enumerate(neg\_word):  plt.text(y, x, str(y), horizontalalignment='left')  plt.xlabel('评论数')  plt.title('贬义词评论词统计')  plt.show()    图5  【步骤二】  热力图  通过热力图，主要是来看耳机不同的方面的相关系数。  *# 热力图*  sns.heatmap(df\_r\_pivot\_tabel.corr(), annot=True)  plt.title('不同主题相关系数热力图')  plt.show()  **四、分析结果**  【步骤一】    图6  可以看到用户多集中在价格，配置和音质上评论，其他方面比较多评论。  【步骤二】    图7    图8  【步骤三】    图9  显然，舒适感，和其他的方面都成负相关的关系，而其他功能上基本都呈正相关关系。    **五、项目结论**  由可视化得到的几个图中来看，用户对耳机的评论大多是关于价格、配置和功能这些方面，由图6可以看出，用户对价格、配置和功能的评价都由较高的好评。但从中评的数据来看，有一部分用户在使用耳机的过程中也遇到了一些问题，给这部分用户带来了一些不好的体验。由图9可以得出，舒适感，和其他的方面都成负相关的关系，而其他功能上基本都呈正相关关系。 | | |
| 项目总结 | 通过《大数据处理分析》这门课我们了解到了用hadoop去处理海量数据，同时在配置hadoop集群环境以及测试中有了更深的了解。并且每个人都配置完了集群环境。在调试的过程中，我们安装好了MySQL和hive环境，但是在利用MySQL制作表格的过程中，结果是出现乱码，所以导致表格的输出没有成功。在结课后我们会再多次尝试MySQL表格的输出。 | | |
| 进一步工作展望 | 由于时间和操作不熟练的问题，我们组最终选用python来完成项目数据的可视化分析。在结课之后，我们还会继续学习hadoop，争取早日掌握其使用方法进行大量的数据分析。 | | |
| 参考文献 | 1. 王润华. 基于Hadoop集群的分布式日志分析系统研究[J]. 科技信息, 2009(15):2. 2. 余慧佳, 刘奕群, 张敏,等. 基于大规模日志分析的搜索引擎用户行为分析[J]. 中文信息学报, 2007, 21(1). 3. 刘飞. 基于网络爬虫的用户行为分析[D]. 汕头大学, 2017. 4. 张丽. 电商平台的用户消费行为分析预测模型[D]. 重庆大学, 2018. 5. 陈小芳, 葛晓滨, 马冠骏. 基于数据挖掘的网络购物用户行为分析[J]. 牡丹江师范学院学报：自然科学版, 2016(1):4. | | |