# 毕业实习报告

| | |
|---|---|
| 学 院 名 称 | 计算机科学与工程学院 |
| 专 业 班 级 | 计算机科学与技术 2018 级 2 班 |
| 学 生 姓 名 | 于跃洲 |
| 学 号 | 201801060228 |
| 指 导 教 师 | 赵华 |

二 O 二二年三月

# 面向特定商品的用户购买原因和用途识别

于跃洲 计算机科学与技术 2018 级 2 班

## 实习地点：山东科技大学

## 实习时间：2022 年 2 月 28 日～3 月 13 日

## 实习内容：

本次实习内容主要是针对毕业设计题目《面向特定商品的用户购买原因和用途识别》所需的相关知识点进行学习并进行初步的应用。

在两周的实习过程中，首先对题目进行了分析，确定了思路为结合自然语言处理（NLP）的相关知识原理对特定商品的用户评论进行文本分析并分类，实现对用户购买原因和用途的识别。在此思路的基础上，遴选出了适宜的特定商品。为使数据特征更加鲜明、更富有多样性，初步计划选择三种特定商品，并对每种特定商品分别进行数据挖掘和文本分析。

| 商品名称 | 是否具有主观预测性 | 用途和购买原因多样性 |
|---|---|---|
| 微波炉 | 有，与做饭有关 | 较为多样 |
| 笔记本电脑 | 无 | 较为多样 |
| 化妆品 | 有，与护肤美容有关 | 较为单一 |

从上表可以看出，这三种商品特征鲜明，且相互之间各有相似点与不同点，这有利于使得本毕设的最终分析结果囊括性更加丰富和全面。

在初步选定待分析的特定商品种类之后，进一步的，初步确定了要使用的技术手段，主要有：

1. 爬虫技术用于数据获取

2. 利用 Python-jieba 进行数据初步处理，如清洗掉非中文字符、文本切片、中文词性标注等。

3.利用 Python-pandas 进行文本词频统计、数据导出等

4. 利用 LDA 模型进行主题分析

基于上述待使用的技术手段，到网络公开的学习平台进行了相关知识的学习。如知网相关论文、Blibli 相关公开课程、Python 官方网站的 pypi。通过学习，了解了上述知识点的基本原理、基础使用方法以及适用场景：

1. 爬虫--数据获取

网络爬虫，其主要原理就是模拟浏览器发送网络请求，接收请求响应，是一种按照一定的规则，自动地抓取互联网信息的程序。因此从理论上来说,只要是浏览器(客户端)能做的事情，爬虫都能够做。故该技术十分适合用于获取完成本毕设所需的相关数据。

2.Python-Jieba 库

Jieba 库是一款优秀的 Python 第三方中文分词库。通过学习，了解到了其分词功能主要有三种模式：精确模式，试图将句子最精确地切开，适合文本分析；全模式，把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义；搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。同时，Jieba 库还支持利用 load_userdict()函数指定自己自定义的词典，以便包含 jieba 词库里没有的词。这是由于 jieba 有新词识别能力，但是自行添加新词可以保证更高的正确率。这十分有利于提高在本毕设分析文本的过程中对新型词汇的识别分析准确率。值得指出的是，自定义词典格式和 dict.txt 一样，一个词占一行；每一行分三部分，一部分为词语，另一部分为词频，最后为词性（可省略），用空格隔开。

3.Python-pandas

通过对 pandas 库的学习，了解到了 Pandas 是一个强大的分析结构化数据的工具集，基于 NumPy 构建，提供了高级数据结构和数据操作工具、提供了大量能够快速便捷地处理数据的函数和方法并且提供数据清洗功能，常应用于数据

挖掘和数据分析领域。

4.LDA 主题模型

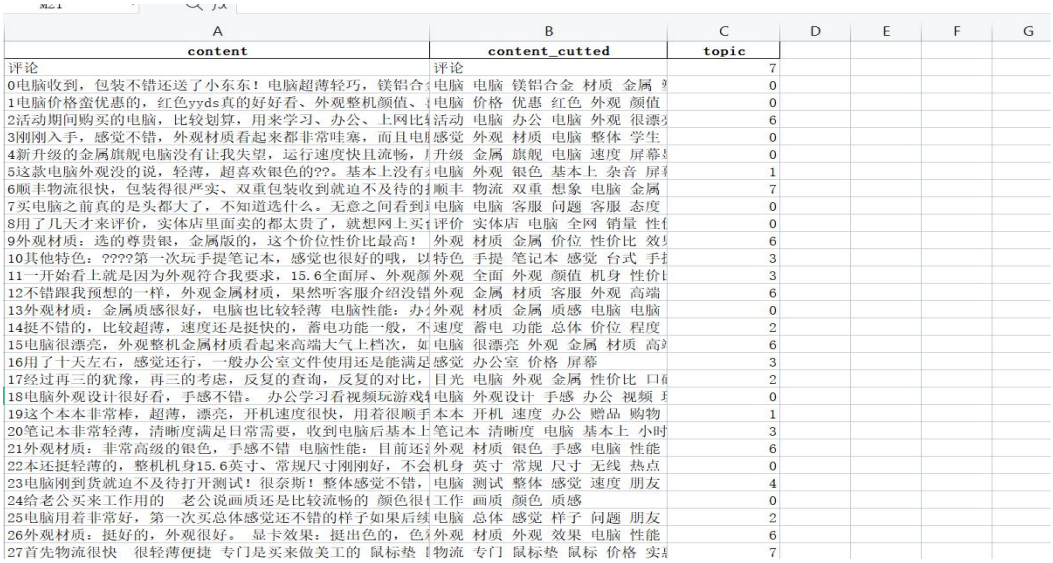
（1）主题模型在自然语言处理等领域是用来在一系列文档中发现抽象主题的一种统计模型。传统判断两个文档相似性的方法是通过查看两个文档共同出现的单词的多少，如 TF（词频）、TF-IDF（词频-逆向文档频率）等，这种方法没有考虑到文字背后的语义关联，例如在两个文档共同出现的单词很少甚至没有，但两个文档是相似的，因此在判断文档相似性时，需要使用主题模型进行语义分析并判断文档相似性。如果一篇文档有多个主题，则一些特定的可代表不同主题的词语会反复的出现，此时，运用主题模型，能够发现文本中使用词语的规律，并且把规律相似的文本联系到一起，以寻求非结构化的文本集中的有用信息。例如微波炉的商品评论文本数据，代表微波炉特征的词语如“加热”“饭菜”“方便”等会频繁地出现在评论中，运用主题模型，把微波炉代表性特征相关的情感描述性词语与应的特征词语联系起来，从而深入了解用户对热水器的关注点及用户对于某一特征的情感倾向。

（2）潜在狄利克雷分配，即 LDA 模型，是一种生成式主题模型，即认为每一篇文档的每一个词都是通过“一定的概率选择了某个主题，并从这个主题中以一定的概率选择了某个词语”。LDA 模型也被称为三层贝叶斯概率模型，包含文档（d）、主题（z）、词（w）三层结构，能够有效对文本进行建模，和传统的空间向量模型（VSM）相比，增加了概率的信息。通过 LDA 主题模型，能够挖掘数据集中的潜在主题，进而分析数据集的集中关注点及其相关特征词。LDA 主题模型是一种无监督的模式，只需要提供训练文档，它就可以自动训练出各种概率，无需任何人工标注过程，节省大量人力及时间。它在文本聚类、主题分析、相似度计算等方面都有广泛的应用，模型的泛化能力较强，不易出现过拟合现象。

（3）LDA 主题模型可以解决多种指代问题，例如：在微波炉的评论中，根据分词的一般规则，经过分词的语句会将“费用”一词单独分割出来，而“费用”是指购买费用，还是维护费用等其他情况，如果简单的进行词频统计及情

感分析，是无法识别的，这种指代不明的问题不能购准确的反应用户情况，运用 LDA 主题模型，可以求得词汇在主题中的概率分布，进而判断“费用”一词属于哪个主题，并求得属于这一主题的概率和同一主题下的其他特征词，从而解决多种指代问题。

在掌握了基本的理论知识后，利用爬虫技术从购物网站对特定商品的用户评论进行了数据爬取，并进行了初步数据清理，对得到的文本利用 jieba 进行了文本词语的切片以及词性的标注，并利用 pandas 完成了各特定商品的词频统计及本地保存。



| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| | content | content_cutted | topic | | | | |
| 0 | 评论 | 评论 | 7 | | | | |
| 1 | 0电脑收到，包装不错还送了小东东！电脑超薄轻巧，镁铝合 | 电脑 电脑 镁铝合金 材质 金属 | 0 | | | | |
| 2 | 1电脑价格蛮优惠的，红色yyds真的好好看、外观整机颜值、 | 电脑 价格 优惠 红色 外观 颜值 | 0 | | | | |
| 3 | 2活动期间购买的电脑，比较划算，用来学习、办公、上网比 | 活动 电脑 办公 电脑 外观 很漂 | 6 | | | | |
| 4 | 3刚刚入手，感觉不错，外观材质看起来都非常哇塞，而且电 | 感觉 外观 材质 电脑 整体 学生 | 0 | | | | |
| 5 | 4新升级的金属旗舰电脑没有让我失望，运行速度快且流畅， | 升级 金属 旗舰 电脑 速度 屏幕 | 0 | | | | |
| 6 | 5这款电脑外观没的说，轻薄，超喜欢银色的??。基本上没有 | 电脑 外观 银色 基本上 杂音 屏 | 1 | | | | |
| 7 | 6顺丰物流很快，包装得很严实、双重包装收到就迫不及待的 | 顺丰 物流 双重 想象 电脑 金属 | 7 | | | | |
| 8 | 7买电脑之前真的是头都大了，不知道选什么。无意之间看到 | 电脑 电脑 客服 问题 客服 态度 | 0 | | | | |
| 9 | 8用了几天才来评价，实体店里面卖的都太贵了，就想网上买 | 评价 实体店 电脑 全网 销量 性 | 0 | | | | |
| 10 | 9外观材质：选的尊贵银，金属版的，这个价位性价比最高！ | 外观 材质 金属 价位 性价比 效 | 6 | | | | |
| 11 | 10其他特色：????第一次玩手提笔记本，感觉也很好的哦，以 | 特色 手提 笔记本 感觉 台式 手 | 3 | | | | |
| 12 | 11一开始看上就是因为外观符合我要求，15.6全面屏、外观颜 | 外观 全面 外观 颜值 机身 性价 | 3 | | | | |
| 13 | 12不错跟我预想的一样，外观金属材质，果然听客服介绍没错 | 外观 金属 材质 客服 外观 高端 | 6 | | | | |
| 14 | 13外观材质：金属质感很好，电脑也比较轻薄 电脑性能：办 | 外观 材质 金属 质感 电脑 电脑 | 0 | | | | |
| 15 | 14挺不错的，比较超薄，速度还是挺快的，蓄电功能一般，不 | 速度 蓄电 功能 总体 价位 程度 | 2 | | | | |
| 16 | 15电脑很漂亮，外观整机金属材质看起来高端大气上档次，如 | 电脑 很漂亮 外观 金属 材质 高 | 6 | | | | |
| 17 | 16用了十天左右，感觉还行，一般办公室文件使用还是能满足 | 感觉 办公室 价格 屏幕 | 3 | | | | |
| 18 | 17经过再三的犹豫，再三的考虑，反复的查询，反复的对比， | 目光 电脑 外观 金属 性价比 口 | 2 | | | | |
| 19 | 18电脑外观设计很好看，手感不错。 办公学习看视频玩游戏 | 电脑 外观设计 手感 办公 视频 | 0 | | | | |
| 20 | 19这个本本非常棒，超薄，漂亮，开机速度很快，用着很顺手 | 本本 开机 速度 办公 赠品 购物 | 1 | | | | |
| 21 | 20笔记本非常轻薄，清晰度满足日常需要，收到电脑后基本上 | 笔记本 清晰度 电脑 基本上 小时 | 3 | | | | |
| 22 | 21外观材质：非常高级的银色，手感不错 电脑性能：目前还 | 外观 材质 银色 手感 电脑 性能 | 6 | | | | |
| 23 | 22本还挺轻薄的，整机机身15.6英寸、常规尺寸刚刚好，不会 | 机身 英寸 常规 尺寸 无线 热点 | 0 | | | | |
| 24 | 23电脑刚到货就迫不及待打开测试！很奈斯！整体感觉不错， | 电脑 测试 整体 感觉 速度 朋友 | 4 | | | | |
| 25 | 24给老公买来工作用的 老公说画质还是比较流畅的 颜色很 | 工作 画质 颜色 质感 | 0 | | | | |
| 26 | 25电脑用着非常好，第一次买总体感觉还不错的样子如果后续 | 电脑 总体 感觉 样子 问题 朋友 | 2 | | | | |
| 27 | 26外观材质：挺好的，外观很好。 显卡效果：挺出色的，色 | 外观 材质 外观 效果 电脑 性能 | 6 | | | | |
| 28 | 27首先物流很快 很轻薄便捷 专门是买来做美工的 鼠标垫 | 物流 专门 鼠标垫 鼠标 价格 实 | 7 | | | | |

进一步的，利用 LDA 模型算法,初步遴选出了一定数量的分好类的文本主题小组，以备进一步的分析。

```
Topic #0:
办公 电脑 游戏 颜值 性价比 软件 视频 玩游戏 电影 评价 问题 方面 有点 系统 画质 屏幕 机身 物流 学生 整体 机子 红色 品牌 价位 体验
Topic #1:
性价比 屏幕 购物 电脑 手感 键盘 款式 质感 本本 声音 体验 大方 银色 外形 美观 分辨率 售后 整体 颜色 尺寸 电池 商家 红色 很漂亮 礼物
Topic #2:
客服 电脑 感觉 问题 态度 物流 正品 小时 总体 想象 很漂亮 实体店 价位 电池 优惠 五星 热情 专业 卡机 价钱 时间 顺丰 试试 礼品 售后
Topic #3:
速度 价格 开机 电脑 实惠 内存 屏幕 画面 物流 功能 颜值 清晰度 网速 发货 杠杠 硬盘 信赖 容量 试试 机器 卡机 机身 玩游戏 色彩 价钱
Topic #4:
笔记本 速度 大气 外观 不卡 朋友 整体 性价比 产品 孩子 时尚 开机 高端 店家 色彩 颜值 工作 显示屏 小巧 玩游戏 价位 银色 活动 键盘 时间
Topic #5:
质量 笔记本电脑 颜色 速度 反应 卖家 电脑 物流 发货 试用 服务态度 孩子 款式 网速 画面 方面 音质 热情 价钱 信赖 卡机 容量 优惠 内存 不卡
Topic #6:
外观 电脑 性能 材质 效果 性价比 特色 很漂亮 外观设计 金属 手感 内存 品质 棒棒 时尚 物流 礼品 音质 不卡 发货 玩游戏 购物 顺丰 售后 朋友
Topic #7:
评论 用户 服务 客服 服务态度 鼠标 物流 鼠标垫 顺丰 商家 时间 下单 地方 键盘 礼物 店家 专业 价钱 学生 显示屏 信赖 尺寸 发货 电脑 软件
```

## 实习体会：

通过这两个周的实习，使我学习到了自然语言处理（NLP）相关的基础知识，并且对毕业设计《面向特定商品的用户购买原因和用途识别》有了较为清晰的实现思路；学习到了如何利用 Python 技术获取到所需数据；学习到了 Python-jieba 库和 Python-pandas 库所具有的功能及基本使用方法和适用场景；了解了 LDA 主题模型的基本原理及其在对文本进行主题提取时的特点和便利性。在汲取了上述理论知识后，将其应用于了本次毕设中。做到了知行合一。

总的来说，本次实习提高了我的专业知识储备量，丰富了我的实际项目经验，加强了我对陌生知识领域的研究能力，清晰了我的后续学业研究方向，使我受益匪浅。