西北工業大學

数据科学的数学方法
大作业

亚马逊商城商品评论数据的研究与挖掘

王嘉利	黄俊淞	彭冠文
2018302278	2018302267	2018302271

2020 年 7 月 19 日

贡　献

王嘉利，黄俊淞，彭冠文 阅读以及分析题目相关信息。

黄俊淞，王嘉利，彭冠文 提取题目核心要素并对完成题目的流程进行了讨论与规划。

王嘉利，彭冠文，黄俊淞 得出了需要用到的技术列表。

彭冠文，王嘉利，黄俊淞 查阅了相关文献并进行了学习研究。

王嘉利，黄俊淞 实现了对文本的分词，清洗文件中出现的符号以及停用词并且完成了词形还原

王嘉利，彭冠文，黄俊淞 基于分词结果实现了 TextRank 和 TF-IDF 完成了关键词提取并对关键词进行打分

黄俊淞，彭冠文 根据关键词提取的相关结果对文本进行向量化

王嘉利，彭冠文 基于 sklearn K-means 完成了文本聚类并于星级评分进行相关分析

黄俊淞，彭冠文 基于 ARIMA 模型对评论和产品的销售情况进行了时间序列的预测和分析

彭冠文 完成了论文的基本模板构建

王嘉利 完成了论文第一章撰写

王嘉利，黄俊淞，彭冠文 完成了论文第二章撰写

彭冠文 完成了论文第三章撰写

黄俊淞 完成了论文第四章撰写

王嘉利 完成了论文摘要的撰写

王嘉利，黄俊淞 完成了论文参考文献和附录的收集整理

摘　要

在本文中，我们旨在使用亚马逊数据中心提供的商品数据来帮助阳光公司进行产品相关的数据分析，在这里，我们总共建立了数个模型。首先建立了文本处理和聚类的模型来分析评分和评论之间的关系。还建立了时间序列分析模型来讨预测文本评论在未来一段时间内的发展。

第一个模型基于 PageRank，TF-IDF，PCA，K-means 等算法。首先，模型对数据集进行预处理，文本的分词和文件的清洗。PageRank,TF-IDF 算法来获取文本中的关键词以及关键词组。之后，通过向量化的方法并且使用 PCA 降维使文本信息成为可以进行聚类的数据化信息。此外，还使用了情感分析的手段判断了评论与情感之间的关系。通过这样做，我们可以找到决定产品成功与否以及影响产品销售的关键因素。

根据 ARIMA 算法，我们构建了第二个用于时间序列分析的模型，旨在分析评论随时间的发展情况以预测产品销售并找出可能成功的产品。我们将评论按月进行统计，并且假设销售量和评论量之间进行正比，建立了自回归整合移动模型（ARIMA）进行预测训练结果表明该模型可以预测销售情况。通过这样的操作，我们得到了预测产品评论走向的方法，并且取得了一定的准确率。

最终，根据上述的结果，我们分析与讨论后向阳光公司市场总监提出建议。

关键词： 自然语言处理，PageRank，TF-IDF，K-means 聚类，PCA，时间序列分析，ARIMA

目 录

第 一 章　背景

世界知名的亚马逊公司创建了在线市场并且为客户提供了对购买进行评级和审查的机会。用户可以通过个人星级评级——从 1 星级到 5 星级来表示他们对产品的满意度，此外，用户也可以通过提交文本信息来进行评论，从而进一步表达对待产品的看法和一些更加具体的意见。用户除了直接针对商品进行评级和评价之外也可以针对其他用户的一些评论进行评级，来表达他们认为这些评论是否对自己购买产品有帮助——从而使在线市场能够筛选出有帮助的评论。

而参与在线市场的产品营销公司则可以通过用户的各种数据数据来洞察他们是否应该参与相关的市场和以及具体的参与时机，这也可以帮助他们对如何设计更受用户欢迎的更成功的产品。

如今，阳光公司正在计划在网上市场推出三种新产品：微波炉，婴儿奶嘴和吹风机。他们聘请我们作为营销团队的顾问。要求我们通过过去用户提供的与其他竞争产品相关的一些评论评级信息来为公司的在线销售战略提供信息，并且确定潜在的重要设计功能，从而增强产品的受欢迎程度。公司曾经使用过数据来为销售战略提供信息，但是并没有使用过特定的组合和类型的数据，他们表示对于这些数据中基于事件的模式非常感兴趣，并且希望能够得到有助于公司制作出成功的产品的方式。阳光数据中心为我们提供了这个项目的三个数据文件：`hair_dryer.tsv`，`microwave.tsv` 和 `pacifier.tsv`。这些数据包括了用户对亚马逊市场中销售的微波炉，婴儿安抚奶嘴和吹风机在一定时间段内的评级和评论信息，希望我们通过这些数据，使用数据科学的相关方法，利用星级评级，评论和评论评级之间的关系和衡量标准等解决一些具体问题或实现一些具体需求：

1. 阳光公司的三种产品一旦在在线市场上线，希望能够根据评级和评论确定出最具信息量的数据衡量标准帮助公司跟踪
2. 在每个数据集中识别并讨论基于时间的度量和模式，这些度量和模式可以表明产品在在线市场的声誉上升或下降
3. 确定基于文本的衡量标准与基于评级的衡量标准的组合，从而能够更好的指示潜在的成功产品或失败产品
4. 具体的星级评分能够带来怎样的影响，是否会引发更多的评论，客户在看到低星级评级后是否会撰写某种类型的评论
5. 基于文本的评论的具体质量描述来判断其是否与评级等级息息相关

第 二 章　方法

基于所提供的数据集，我们将每一条信息分为如下的维度：

文本 每一条评论的文字中都包含了丰富的语义信息。虽然文本信息是非结构化，非数字化的，但现在数据科学领域内有着大量成熟的自然语言处理技术，通过一系列算法的变换后，可以得到包含语义信息的数值化数据。

时间 商品的声誉并非是静态的，而是随着时间变化的。数据集中给出了每一条评论的时间戳，通过结合其他数据维度，可以获得数据之间的时序关联。

评分 评分是用户对产品评价最基本的数值度量，平均评分也是最为容易取得的可靠指标。

评价数 出于合理的推断，我们假设一个商品的评价量 C 总是占总销售量 S 的一个固定比例，即满足 $C/S = \text{const}$。又因为销售量和产品营收成直接的正相关关系，所以评价数也可以作为一个重要的指标。

在经过基础的数据处理后，我们对不同维度的指标做交叉分析。基本的关系图如图 2-1 所示。

2.1 文本处理

本文拟采用分词，关键字提取，文本向量化等方法对文本进行处理，以得到便于计算的向量值。基本流程图见图 2-2。

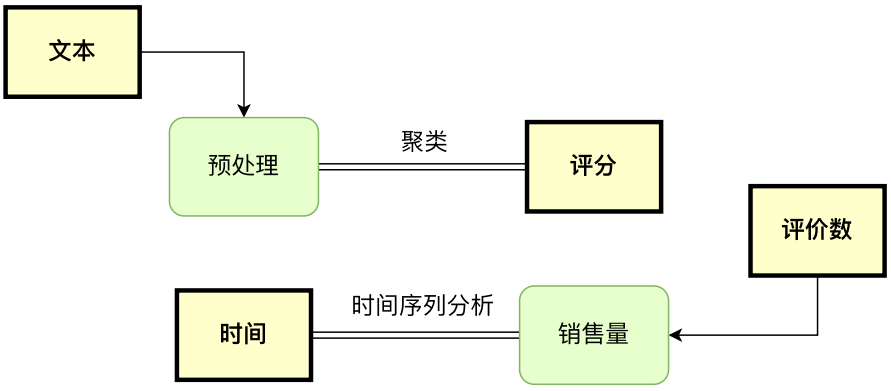

图 2-1　数据关系

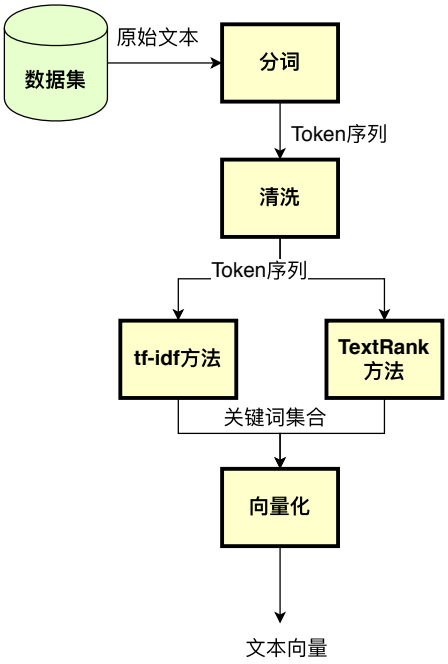

图 2-2　文本处理流程

2.1.1 分词（Tokenization）

数据集中存在的文本是英文文本，可以使用一些英语的自然语言处理库来进行分词处理经过分析与讨论，我们决定使用自然语言处理库 NLTK[1] 的 `tokenize` 函数来进行处理。该函数和的分词方法和 Penn TreeBank 语料库的分词方法一致，TreeBank 标记生成器使用正则表达式对文本进行标记化。该过程由调用的方法 `word_tokenize()`。假设文本已经被分割成句子，则使用 `sent_tokenize()` 来进行操作

该标记器执行以下步骤:

1. 分割标准收缩，例如将 `don't` 转化为 `do n't`
2. 将大多数标点符号视为单独的标记进行分割
3. 分隔逗号和单引号以及句末的空格
4. 若出现了行末的单独的句点，将其分割开

2.1.2 文本清洗（Text Preprocessing）

数据集中的用户评论文本是非结构化的复杂数据，有许多干扰因素。在利用自然语言算法进行进一步处理之前，需要对文本进行清洗操作。

i me my myself we our ours ourselves you you're you've you'll you'd your yours yourself yourselves he him his himself she she's her hers herself it it's its itself they them their theirs themselves what which who whom this that that'll these those am is are was were be been being have has had having do does did doing a an the and but if or because as until while of at by for with about against between into through during before after above below to from up down in out on off over under again further then once here there when where why how all any both each few more most other some such no nor not only own same so than too very s t can will just don don't should should've now d ll m o re ve y ain aren aren't couldn couldn't didn didn't doesn doesn't hadn hadn't hasn hasn't haven haven't isn isn't ma mightn mightn't mustn mustn't needn needn't shan shan't shouldn shouldn't wasn wasn't weren weren't won won't wouldn wouldn't

图 2-3　停用词表

首先要去除非文本字符和标记。在文档中含有用于在网页上渲染格式的 HTML 标记 `<br/>`，以及用于显示视频的 `[[img:]]` 标记等。这些字符对于语义分析和自然语言处理算法都没有价值，应当予以去除。

接下来考虑词汇的同一性的问题。英语单词在句首时需要大写首字母，但是这样的变化并不带有语义信息；同时英语属于拉丁语系，单词具有不同的屈折变化（Inflection)。不同的屈折变化语义相近，应该做合并处理。为了方便计算机处理，我们对单词做标准化（Normalization）处理，将大写字母转换为小写，并使用 NLTK 库的词性还原功能进行处理。这样标准化之后的词汇可以直接用字符串比较来判断同一性。

最后，一个停用词表（Stopwords List）给出了所有应该被去掉的虚词和信息量小的实词。所有在该列表中的单词都被从文档中移除。本文使用了来自 NLTK 库的英语停用词表，一共包含 179 个单词，见图 2-3。

2.1.3 关键词提取（Keyword Extraction）

常见的关键词提取方法有 tf-idf [2]，PageRank,Topic-model 等算法。本文使用 tf-idf 算法和 PageRank 算法的衍生算法 TextRank 算法 [3] 来进行关键词提取的操作。

(1) if-idf 方法

tf-idf 是一种用于信息检索与数据挖掘的常用加权技术。tf 是词频 (Term Frequency)，idf 是逆文本频率指数 (Inverse Document Frequency)。它是一种统计方法，经常用来评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。tf-idf 的主要思想是：如果

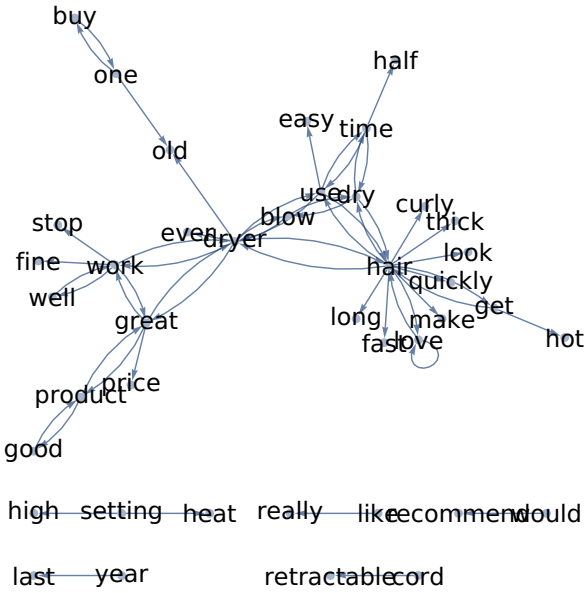

图 2-4　关键词图 V 的一部分。选取了频数最大的 20 个元素，权值小于 200 的边被隐去。可以看出常见的词语组合如“last year”“buy one”等。

某个词或短语在一篇文章中出现的频率 tf 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。其中 tf 的计算方法为:

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2.1.1}$$

idf 的计算方法为:

$$\mathrm{idf_i} = \lg \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{2.1.2}$$

分别计算出相应的 tf 和 idf 的值之后，将他们求积就可以得到对应的 tf-idf 值:

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i \tag{2.1.3}$$

(2) TextRank 方法

PageRank 设计之初用于 Google 的网页排名。其具体公式如下:

$$S(V_i) = (1-d) + d \cdot \sum_{j \in \mathrm{In}(V_i)} \frac{1}{|\mathrm{Out}(V_j)|} S(V_j) \tag{2.1.4}$$

该公式中，V_i 表示某个网页，V_j 表示链接到 V_i 的网页，$S(V_i)$ 表示网页 V_i 的 PR 值，$In(V_i)$ 表示网页 Vi 的所有入链的集合,$Out(V_j)$

我们使用了 PageRank 算法的衍生算法 TextRank，与 PageRank 算法相比，它多了一个权重，用来表示两个节点之间的边连接有不同的重要程度，

迭代公式如下:

$$WS\left(V_i\right)=(1-d)+d\cdot\sum_{V_j\in\mathrm{In}(V_i)}\frac{w_{ji}}{\sum_{V_k\in\mathrm{Out}(V_j)}w_{jk}}WS\left(V_j\right) \tag{2.1.5}$$

该算法的具体过程如下。

1. 构建候选关键词图 $G_K=(V,E)$，其中 V 为节点集, 由上一步生成的候选关键词组成，然后采用共现关系构造任两点之间的边，两个节点之间存在边仅当它们对应的词汇在长度为 w 的窗口中共现，w 表示窗口大小，即最多共现 w 个单词。关键词图的一部分见图 2-4 。
2. 根据式 2.1.5 ，迭代传播各节点的权重，直至收敛。
3. 对节点权重进行倒序排序，从而得到最重要的 m' 个单词，构成集合 $K'=\{K_j\}$

2.1.4 文本向量化

基于提取的关键词，我们可以针对文档做文本向量化。文本向量化将由标记组成的文本 $T_n=\{t_1,t_2,\ldots,t_n\}$ 转换为向量 $W_T=\{k_1,k_2,\ldots,k_m\}$，其中 n 为文本的长度，m 为向量的长度。为了找出关键词集合，首先需要选定一组序数为 m 的关键词集合 $K=\{K_i\}$，向量分量 k_i 为对应关键词在文章 T 中的 tf-idf 权重。

$$k_i=\mathrm{tfidf}_T(K_i) \tag{2.1.6}$$

集合 K 由 TextRank 方法选出的关键字集合 K'，再由 tf-idf 方法做反向筛选得出。

$$K=\{K_i|K_i\in K',K_i\in\text{top 1000 of }K'_{\text{tfidf}}\} \tag{2.1.7}$$

2.2 评分与文本的关联聚类

通过前面的步骤我们得到了一个数千维的向量 V, 为了进行接下来的聚类操作，我们决定先将这个高维向量降维处理，结合课上所学内容，我们分析讨论之后决定使用主成分分析法（PCA）来进行降维操作。

PCA（Principal components analysis）是一种统计分析、简化数据集的方法。其具体实现步骤如下:

1. 我们首先对特征进行归一化处理，以便于继续进行后面的操作

$$\begin{cases} \mu_j = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} x_j^{(i)} \\ x_j^{(i)} = x_j^{(i)} - \mu_j \\ \sigma_j^2 = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}\left(x_j^{(i)}\right)^2 \\ x_j^{(i)} = \dfrac{x_j^{(i)}}{\sigma_j} \end{cases} \tag{2.2.8}$$

2. 计算协方差矩阵。

$$C = \frac{1}{m}X^T X \tag{2.2.9}$$

3. 我们计算了向量的协方差矩阵的特征向量并按照特征大从大到小排序。

4. 提取出特征向量矩阵的前 k 列。

$$U = \begin{pmatrix} \boldsymbol{u}^{(1)} & \boldsymbol{u}^{(2)} & \cdots & \boldsymbol{u}^{(k)} \end{pmatrix} \tag{2.2.10}$$

5. 通过矩阵乘法计算得到新的特征 Z。

$$Z = XU = \begin{pmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \\ \vdots \\ \boldsymbol{x}^{(m)} \end{pmatrix} \begin{pmatrix} \boldsymbol{u}^{(1)} & \boldsymbol{u}^{(2)} & \cdots & \boldsymbol{u}^{(k)} \end{pmatrix} \tag{2.2.11}$$

使用 PCA 降维之后我们成功的将原本的高维向量 V 降维到 d 维（$d = 100$），使向量的维度降低到可以操作的程度了，在这之后，我们利用了 K-means 聚类算法进行了文本聚类的操作，以下是基于 python sklearn 库的 K-means 算法实现方法：

1. 我们首先随机选择了 k（$k = 10$）个中心

2. 随后遍历所有样本，把样本划分到距离最近的一个中心

3. 划分之后就有 K 个簇，计算每个簇的平均值作为新的质心

4. 不断重复上面的步骤，直到满足停止条件

我们设定的停止条件为该算法的默认停止条件：即聚类中心不再发生变化。当算法运行到满足停止条件的时候我们得到了一幅饱含信息的图用以进行后面的分析

2.3 评价量的时间序列分析

我们将评论按月进行统计得到月评论量 x_i，并且假设销售量 X_i 和评论量 x_i 之间成正比：

$$X_i \propto x_i \tag{2.3.12}$$

建立自回归整合移动（ARIMA）模型 [4], 对未来的销售量 w_t 进行预测。为了得到平稳的序列，对月评论量进行差分处理。

$$\begin{cases} w_t &= \Delta^d x_t \\ w_t &= \phi_i w_{t-1} + \cdots + \phi_p w_{t-p} + \delta + t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q} \end{cases} \tag{2.3.13}$$

其中 $\phi_i w_{t-1} + \cdots + \phi_p w_{t-p}$ 对过去值进行自回归，Δ^d 为 d 阶差分运算，考虑历史数据对未来的影响, 在与过去的白噪音移动平均 $\theta_1 u_{t-1} + \cdots + \theta_q u_{t-q}$ 进行组合得到 ARIMA 模型的参数通过分析自相关图和偏自相关图确定模型参数。

第 三 章　结果

3.1 评分与文本的关联聚类

在文本向量化的基础上，对文本进行聚类分析图 3-1 ，并与评论评星等级进行对比图 3-2。

并且文本进行情感分析，计算和评分等级之间计算 Pearson 相关系数为 0.39，因此可以认为评分等级和评论的情感具有较强的相关性。

3.2 文档分类

我们基于支持向量机对文档向量进行分类，预测评论的平行等级。抽取 7000 组数据作为训练集，将剩余的数据作为测试集，可以从图 3-3 中看到，产品评级等级以 5 星和 4 星为主，总体上评星等级较高，预测结果正确率为 0.53。

3.3 时间序列分析预测评论量

将数据进行整理得到月评论量，为了得到平稳序列，对月评论量据进行 2 阶差分图 3-4，通过 ADF 检验认为得到序列是平稳的。通过自相关图和偏自相关图图 3-5 确定 ARIMA 的参数为

$$p = 12, d = 2, q = 0 \tag{3.3.1}$$

对 ARIMA 进行训练，将月评论量的后 20 个数据作为测试集，并与真实评论量做对比，从图 3-6 中可以看出预测结果在短期时预测结果与真实值接近，在较远的时间预测的趋势和真实评论量趋势一致.

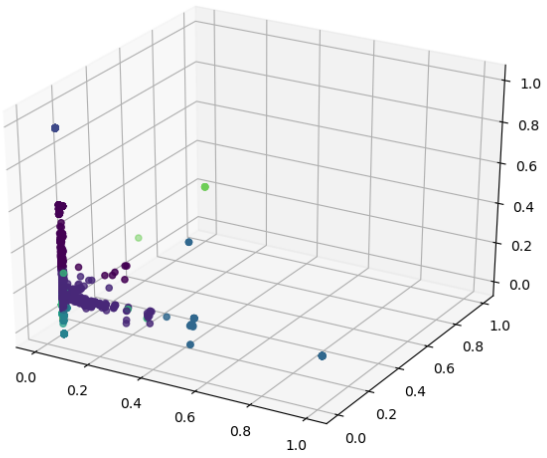

图 3-1　K-means 聚类

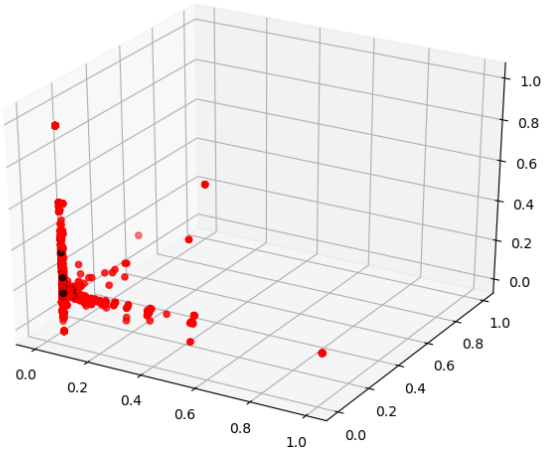

图 3-2　评分等级分类，黑色表示低评分，红色表示高评分

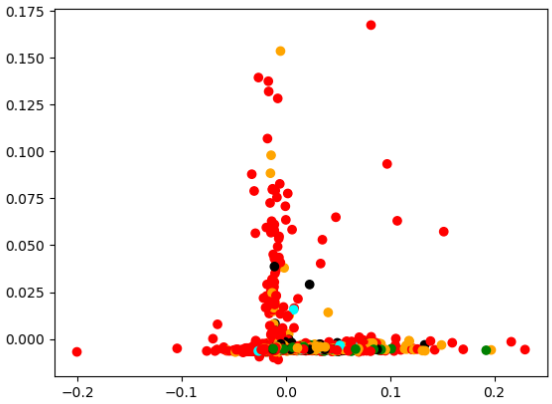

图 3-3　评论评星等级，红色代表 5 星，橙色代表 4 星，绿色代表 3 星，蓝色代表 2 星，黑色代表 1 星

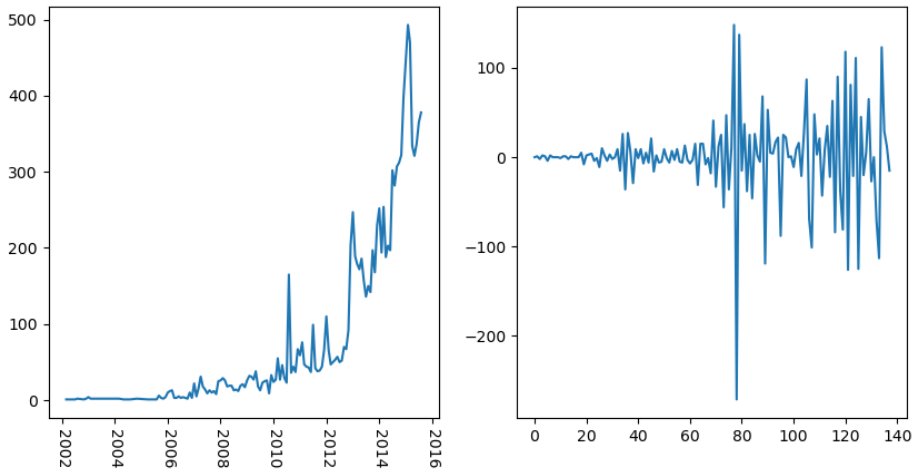

图 3-4　月销售数据和 2 阶差分结果

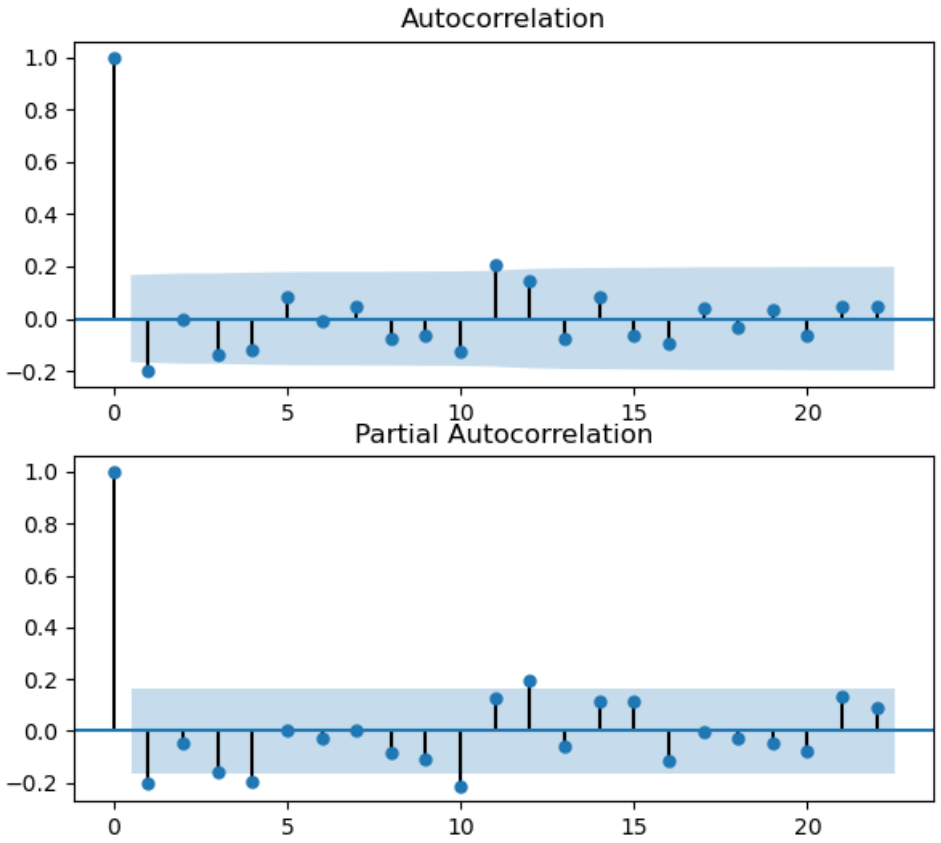

图 3-5　ACF PACF

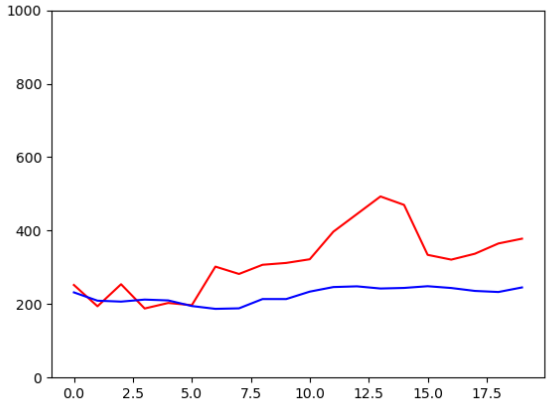

图 3-6　ANIMA 时间序列分析结果分析

第 四 章　讨论

本文基于亚马逊的在线市场为背景，通过使用一些数据科学的数学方法来完成了对阳光公司相关产品的一些预测与分析，在实践过程中，我们利用各种方法，采用了不同的模型来处理来自亚马逊数据中心的用户评论数据集。首先，我们进行了对文件的预处理，获得了干净的而简洁的数据集，然后这些数据来获得评论中的关键词，之后对文本信息进行向量化并且使用 PCA 进行降维处理，然后对完成了文本聚类的操作并且还根据时间序列和评论的相关信息得到了一些更多的信息。在进行时间序列分析时，我们得到的预测曲线与实际情况基本相符，但是缺少一个峰值，经过讨论，我们认为可能是数据样本量不够大导致的。

最终，我们得到了一些想要的结果并且完成了阳光公司的提出的问题，通过对这些结果进行分析，我们能够向阳光公司提出一些建议。

此外，本次实践在对于文档分类的操作上还有可以提升的空间，目前我们达到了 0.53 准确率，但是如果使用一些其他的算法，可能还有提升的空间。

参考文献

[1] Loper E, Bird S. Nltk: the natural language toolkit[J]. arXiv preprint cs/0205028, 2002.

[2] Ramos J, et al. Using tf-idf to determine word relevance in document queries[C]// Proceedings of the first instructional conference on machine learning: volume 242. New Jersey, USA, 2003: 133-142.

[3] Mihalcea R, Tarau P. Textrank: Bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. 2004: 404-411.

[4] Contreras J, Espinola R, Nogales F J, et al. Arima models to predict next-day electricity prices[J]. IEEE transactions on power systems, 2003, 18(3):1014-1020.

[5] Shen S, Wang Z, Zhang J, et al. LaTeX-template-for-npu-thesis[Z]. 2016.

[6] 王斌. 信息检索导论[M]. 北京: 人民邮电出版社, 2010.

[7] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python [J]. Journal of Machine Learning Research, 2011, 12:2825-2830.

[8] Nielsen F Å. A new anew: Evaluation of a word list for sentiment analysis in microblogs [J]. arXiv preprint arXiv:1103.2903, 2011.