**(封面A3牛皮纸，内容与该页相同，封面内保留该页内容)**

**(页面设置：A4纸打印，版面上空2.5cm，下空2.0cm,左空2.2cm，右空2.2cm,装订线0.5cm)**

**山 东 科 技 大 学**

**毕 业 设 计（论 文）**

**说 明 书**

**设计题目：** （宋体正文小二号字）

**院 系： 专业： 班级：**

**设 计 人：**

**2020年 月 日**

|  |
| --- |
| **毕业设计任务书**  **专业 班 学生**   1. 毕业设计题目： 2. 设 计 专 题：   三、设计原始资料：      四、设计应解决下列各主要问题：        五、设计说明书应附有下列图纸：      六、命题发出日期： 年 月 日 设计应完成日期： 年 月 日  设计指导人（签章）：  教研室主任（签章）：  院 长（签章）： |

|  |
| --- |
| **特 约 评 阅 人 意 见**    评阅人（签章）：  评阅人职称：  日 期： 年 月 日 |

|  |
| --- |
| **指导教师对毕业设计的评语**    指导教师（签章）：  日 期： 年 月 日 |

|  |
| --- |
| **答辩委员会鉴定意见**    答辩成绩（等次及分数）：  **鉴定意见：**  答辩委员会主任（签章）：  答辩委员会副主任（签章）：  日 期： 年 月 日 |

**摘 要**

随着经济的发展，网络中涌现出大量的文本数据，这些文本数据的特征与其所属领域有着紧密的联系，于是产生许多文本分析任务。目前传统文本分析任务多数在同一领域，即对于某一领域已有标签的文本数据进行模型训练，将训练出来的模型同样应用于本领域。但是在当前网络环境下，存在着各类领域，其中许多领域的数据是没有标签的，使用人工标注的方式非常的耗时耗力，因此便产生了对面向文本处理的迁移学习任务方面的需求。然而，源领域与目标领域文本之间的特征往往是有较大差别的，若想使源领域中训练得出的模型也能够在目标领域中使用，关键是做好两个领域间的特征对齐任务。

跨领域的文本特征对齐任务的相关流程主要有文本数据采集、数据预处理、特征提取、特征对齐和模型检验五个过程。本课题中以电商领域为源领域，以新闻领域为目标领域，分别对这五个过程开展研究。首先使用采集工具获取两个领域中用户评论文本的数据，随后使用jieba分词工具进行文本分割与停用词去除，之后使用词频-逆文档频率(Term Frequency–Inverse Document Frequency, TF-IDF)统计法统计出每个领域中文本特征词，以保证最终训练出来的模型能够识别出某一文本样例中包含的特征词。在文本特征对齐的工作中，本文提出一种基于Word2vec工具的特征对齐算法来进行两个领域之间的特征对齐，此算法要求除枢纽特征之外，对目标领域中的每个非枢纽特征，在源领域的特征词中使用Word2vec工具查找出与之相似度最高的特征，从而建立两个领域之间的特征对应关系，这样在最后做预测时，通过将源领域中相似特征在目标领域中进行替换，最后在源领域中训练出一个基于支持向量机(SVM)的分类器，将其迁移到目标领域，测试文本分类效果。

结果显示，通过本文提出的特征对齐算法，分类器在迁移完成后的分类准确率

达

到

**关键词：**文本特征对齐、jieba分词技术、TF-IDF算法、Word2vec工具、支持向量机(SVM)

**Abstract**

In recent years, with the development of modern science and technology, various kinds of social media (such as news websites, social websites, e-commerce websites, etc.) have risen rapidly, which makes a large number of text data emerge in the network. The structure of these text data is often related to the characteristics of their fields. Therefore, cross field transfer learning is developed in the fields of natural language processing and machine learning. At present, the goal of traditional text analysis is mostly in the same field, that is, learning a model for the existing tagged text data in a certain field, and applying the trained model to this field. However, in the current network environment, there are various fields and they are also full of complexity, so too many data are unlabeled, the use of manual annotation will waste a lot of human and material resources, at the same time, there will be a lot of subjectivity in the annotation results, so there is an urgent need for some cross domain text annotation algorithms.

The related processes of cross domain text feature alignment algorithm include text data collection, text preprocessing, feature extraction, feature alignment and model verification. In this thesis, I take the feature alignment process of e-commerce and news as an example, and study the five processes respectively, and the algorithm of the last three steps. After the completion of data collection, I first use the Jieba word segmentation tool to segment the text and remove the stop words, and then I use the TF-IDF statistical method to extract the text features, and then a feature alignment algorithm based on word2vec tools is used for feature alignment between two fields. finally, support vector machine (SVM) model is established to test the text classification effect in the target field.

The results show that the accuracy of the classifier after migration is significantly improved by the feature alignment algorithm proposed in this thesis.

**Keywords:** Text feature alignment, jieba, TF-IDF algorithm, word2ve, support vector machine (SVM)

**目 录**

[1 绪 论 1](#_Toc41032142)

[1.1 课题的提出 1](#_Toc41032143)

[1.2 文本特征对齐问题的研究现状 2](#_Toc41032144)

[1.3 课题研究的目的 4](#_Toc41032145)

[2 相关概念介绍 5](#_Toc41032146)

[2.1 文本分类 5](#_Toc41032147)

[2.2 情感分类 5](#_Toc41032148)

[2.3 迁移学习 6](#_Toc41032149)

[2.3.1 基于样本的迁移学习 6](#_Toc41032150)

[2.3.2 基于基于相关性的迁移学习 6](#_Toc41032151)

[2.3.3 基于模型的迁移学习 7](#_Toc41032152)

[2.3.4 基于特征的迁移学习 7](#_Toc41032153)

[2.4 文本特征与特征选择 7](#_Toc41032154)

[2.4.1 卡方统计(CHI)法 7](#_Toc41032155)

[2.4.2 信息增益(Information Gain, IG)法 8](#_Toc41032156)

[2.4.3 互信息(Mutual Information, MI)方法 8](#_Toc41032157)

[2.4.4 词频-逆文档频率(Term Frequency–Inverse Document Frequency, TF-IDF)统计法 9](#_Toc41032158)

[2.5 文本特征对齐 10](#_Toc41032159)

[2.6 分类器 10](#_Toc41032160)

[2.6.1 支持向量机(SVM) 11](#_Toc41032161)

[2.6.2 朴素贝叶斯(Naïve Bayesian Classifier) 12](#_Toc41032162)

[2.6.3 决策树(Decision Tree) 13](#_Toc41032163)

[2.6.4 k近邻(k-nearest neighbor) 14](#_Toc41032164)

[2.7 本章小结 17](#_Toc41032165)

[3 文本特征对齐任务及数据预处理 17](#_Toc41032166)

[3.1 文本特征对齐任务 17](#_Toc41032167)

[3.1.1 任务介绍 17](#_Toc41032168)

[3.1.2 工作流程 18](#_Toc41032169)

[3.1.3 评价指标 19](#_Toc41032170)

[3.2 文本搜集 20](#_Toc41032171)

[3.2.1 源领域 20](#_Toc41032172)

[3.2.2 目标领域 20](#_Toc41032173)

[3.3 数据预处理 21](#_Toc41032174)

[3.3.1 数据认知 21](#_Toc41032175)

[3.3.2 数据预处理 21](#_Toc41032176)

[3.4 本章小结 22](#_Toc41032177)

[4 特征提取 23](#_Toc41032178)

[5 特征对齐 23](#_Toc41032179)

[6 分类器构造与模型评估 23](#_Toc41032180)

[7 总结与展望 23](#_Toc41032181)

[参考文献 25](#_Toc41032182)

# 1 绪 论

## 1.1 课题的提出

近年来，随着当代科学技术的发展，各种社会媒体（如：新闻网站，社交网站，电商网站等）迅速崛起，使得人们可以通过越来越多的方式在网络上发表自己的观点，致使当前国内外网络环境中涌现出大量的文本数据。这些文本数据的结构与其所属领域特点往往有着较大的联系，有些词汇甚至可以成为某些领域的代名词，例如：当我们看到“好评”、“物美价廉”等词语时，我们会想到这可能是对于某一商品的评价；当我们看到“吓人”、“振奋人心”等词语时，我们会想到这可能是对于某一消息或新闻的评价。于是一些情感分类算法在机器学习领域应运而生，但是这些情感分类算法的分析目标多数是在同一领域，即对于某一领域的已有标签的文本数据进行机器学习，将训练出来的模型同样应用于本领域，从而完成对未分类的文本的分类任务。但是在当前网络环境下，各类领域及其繁多复杂，而更多的领域的文本数据是没有标签的，同时使用人工标注的方式会浪费极大的人力物力，并且标注结果也会存在很大的主观性，导致标注结果的准确性极大的降低，因此在文本分析中亟需一些可实现跨领域文本标注的算法[1]，来对已有文本进行情感分类。

笔者认为，目前国内经济发展水平呈上升态势，人民群众正在不断谋取更高的物质生活水平与精神生活水平，于是网上购物与新闻知识获取的向往日趋强烈，使得电商领域与新闻领域成为全体网民乐于聚集之地，这两个领域所涵盖的人群数量大且范围广，因此在两个领域中留下了的文本信息较为全面，对这两个领域做迁移学习将有更大的研究价值，例如国家可以通过文本情感分析结果对社会舆论信息进行监控；对于包括新闻网站在内的大多数网站来说，可以对用户评论内容进行关键词提取进而对不符合要求的敏感词进行屏蔽；电商平台可以利用用户评论信息进行产品优化或修改营销计划；一些音视频媒体也可以根据用户对某些影视剧或某些歌曲的评论情况对其他用户进行择优推荐等等。目前面向迁移学习的文本情感分类并没有成文的步骤，笔者在阅读大量文献之后，总结得出做文本情感分类方面的迁移学习应当有如下4个步骤：1、文本预处理，2、文本特征提取，3、特征对齐，4、模型检验。显然，第2与第3步应当是整个过程的重中之重，对于基于特征的迁移学习而言，源域与目标域之间的枢纽特征的匹配对于在两个领域中进行机器学习的效果是至关重要的，因此研究好如何将两个领域中的非枢纽特征更好地转化为枢纽特征，以及如何将两个领域中的枢纽特征做最大程度上的对齐，对于提升迁移学习的准确率是至关重要的，这是选取本课题的重要意义之所在。

本文中，笔者首先对当今时代下跨领域文本特征对齐算法进行全面剖析，而后选取对本课题最为契合的算法，以电商领域为源域，以新闻领域为目标域，完成两个领域之间的迁移学习。

## 1.2 文本特征对齐问题的研究现状

目前网络环境中，各类文本信息有着多种多样的形式，它们所表达的情感倾向也相差甚大，例如在电商领域与新闻领域中，用户们会发表许多的关于产品或者时政的评论及意见，这些评论及意见通常含有用户的情感极性。Turney[2]等人在2002年第一次将文本情感分类的任务当做学术问题公开提出，之后再行业内引发了众多学者的研究。

目前常用的基于统计的特征选择方法有信息增益方法(IG)、互信息方法(MI)、卡方统计(CHI)法、文档频率方法(DF)等[3-5]。2010年，在总结已有特征选择方法的基础上，杨凯峰[6]等一种新的基于文档频率(DF)的特征选择方法，此方法主要用于中文文本分类，需要先对每篇文档进行分词处理。他们使用 ICTCLAS(Institute of Computing Technology,Chinese Lexical Analysis System)系统完成文档的分词及词性标注，使用特征词在一个类别中出现的文档数来表示这个特征词与该类别的相关度，他们使用复旦大学整理的语料库进行分类实验，结果显示，在进行特征选择时，补充部分高词频特征词可以提高分类的召回率和准确率。在向量空间模型中，将文本表示单元（特征词）转换为向量通常使用的权重计算方法是TF-IDF(term frequency-inverse document frequency)方法，在一定程度上，这个方法能有效地表示一个特征词在文本中区分文本属性的重要程度，但其理论依据存在不足。2013年路永和[7]等提出了改进TF-IDF算法的文本特征项权值计算方法，改进的核心思想是：每个特征词对每个类的区分能力不同，其重要性用特征选择评估函数反映，同时该词集中分布的类就是其所属的类，即这个特征词要具有这个类的特征性。为避免特征词只集中出现在某篇文档中，要求该特征词在该类别文档中应均匀分布。随后，曹鲁慧[8]等提出一种基于深度学习的中文文本特征提取与分类方法，她们使用卷积神经网络(CNN)进行文本分类，将每个样本用128维的向量进行表示，大大减少特征的维度，加快分类器的训练速度，进而提高分类的准确率。卢晨阳[9]等提出一种基于语义结构的迁移学习文本特征对齐算法，他们使用wordnet模型完成领域独立特征的对齐。他们考虑到，传统的枢纽特征抽取和对齐一般只考虑到句子的结构和词性等信息，而忽略词语特征和类别信息存在一定的相关性，因此在抽取特征时应考虑类别影响。以情感分类为例，在抽取特征训练模型时，首先将源域和目标域的正类样本(积极)和负类样本(消极)分开，将不同词性的词语抽取出来，其中，在情感分类中起主要影响的是名词、动词、形容词、副词，然后对正类样本和负类样本分别抽取名词、动词、形容词和副词，原始的源域和目标域的2份数据变成根据样本类别和词性划分的16 份数据。最后 使用16 份数据作为词向量模型GloVe 模型的输入来进行训练。

另外，学者们对文本特征对齐算法的研究也在不断深入，2006年，John Blitzer[10]等提出一种结构一致学习算法(Structure Correspondence Learning, SCL)，这种算法可以自动归纳来自不同领域的特征之间的对应关系，SCL的核心思想是通过建立不同领域特征与枢纽特征的相关性来识别特征之间的对应关系。SCL的第一步是在两个域的未标记数据上定义一组枢纽特征。然后，使用这些枢纽特征来学习从两个域的原始特征空间到共享的低维实值特征空间的映射。在这个新的空间中，如果内积很高，则表示对齐效果较好。在有监督的训练任务中，同时使用源域中的转换特征和原始特征。在有监督的测试任务中，则同时使用目标域的转换特征和原始特征。如果学习到的映射比较好，那么在源域上学习到的分类器也将在目标域上有效。2013年，孟佳娜[11]等提出一种基于特征变换的跨领域产品评论倾向性分析方法，她们使用一种特征变换算法(Feature Transform Algorithm,FTrA)，首先计算源领域和目标领域的领域独立词，然后，求得源领域的领域依赖词与每个领域独立词之间的关联度值，与某个领域独立词关联度最高的特征，与其相关性越高。同时，求得目标领域的领域依赖词与每个领域独立词之间的关联度值，与某个领域独立词关联度值越高的特征，与其相关性越高。在此基础上，Sinno Jialin Pan[12]等提出了一种谱特征对齐 (Spectral Feature Alignment,SFA) 算法。SFA以一些领域无关词为桥梁，构造了一个二部图，来模拟领域特定词和领域无关词之间的共现关系。其思想是：如果两个特定于域的词与图中更常见的独立于域的词有关联，它们往往会以更高的概率对齐。类似地，如果两个独立于域的词与图中更常见的特定于域的词有连接，则它们倾向于以更高的概率排列在一起。他们采用基于图谱理论[13]的谱聚类算法，在二部图的基础上，将特定于领域的和与领域无关的词联合成一组特征聚类。这样，集群就可以用来减少两个域的特定于域的词之间的不匹配。最后，他们用这些聚类来表示所有的数据实例，并基于新的表示训练情感分类器。与结构一致学习(SCL)[10]等跨域情感分类算法不同，他们提出的SFA算法可以通过在二部图上的联合对齐，充分利用领域无关词和领域特定词之间的关系，学习更紧凑、更有意义的表示图的底层。通过实验表明，在跨领域情感分类的准确性方面，SFA确实比SCL等具有更好的性能。但是Yuhong Zhang[14]等认为，SCL与SFA这类的算法尝试提取出域无关主题词来构造子空间，并在子空间中训练分类器，这样的话，目标域中的域相关特征会被映射为源域的特征，然而在实际应用当中，这些方法无法解决在不同域中具有不同情感的特征词的问题，因为一个主题仅仅能够表达出一种情感，这些特征就会在源域中被错误地训练。针对这一问题，Yuhong Zhang等人提出一种面向跨领域情感分类的公共子空间构造算法[14]，此算法的目标是构造出一个更加精确的公共子空间，然后就可以在这个子空间中为目标域训练出分类器。对于有标签的源域S和无标签的目标域T，他们使用分类比例差[15]( Categorical Proportional Difference, CPD)算法计算源域中特征词的情感倾向，然后基于共现关系预测目标域中这些特征的情感取向，根据不同域之间情感取向的一致性构造公共子空间，于是域相关特征将会被过滤掉，最终使所有的域相关特征都会被投影到公共子空间中来解决域失配问题。跨领域的文本特征对齐是做文本情感分析的迁移学习过程中的重要一步，对与文本特征对齐技术与算法方面的研究，时至今日，仍吸引着众多的学者来专研，这也是笔者以此为课题的原因之一。

## 1.3 课题研究的任务

本课题旨在通过研究前人在文本特征对齐方面的优秀经验的基础上，研究出一套适用于本课题研究背景的文本特征对齐算法。

然而，跨领域的文本特征对齐并不是简单的一个算法既可以全盘解决的，特征值的计算，特征对齐方法的实现，分类模型的构建等问题是一脉相连下来的，若要完成文本特征对齐任务或者计算出特征迁移成功率就必须实现这一系列任务中的每一个，因此在本课题中对文本特征对齐任务中的五个步骤（文本数据采集、数据预处理、特征提取、特征对齐、模型检验）均做了细致实验，并在本文中详细描述出来。

下面对本论文的论文结构做出介绍：在第二章中，对文本特征对齐任务的相关概念作出解释，对一些重点算法作出适当解读；在第三章中，前半部分介绍本研究课题中使用的数据来源以及采集计划，后半部分介绍对这些数据的预处理过程；在第四章中，本文首先对所使用的特征提取算法——TF-IDF算法进行细致介绍，随后给出TF-IDF算法的计算程序，并对该程序的运行原理展开介绍，最后给出程序的运行结果；在第五章中，需要进行文本特征对齐的工作，这一步的目的是找出源领域与目标领域中特征之间的匹配关系，这应当是一种一对一的配对关系，有了这个关系，后面训练出来的机器学习模型才能够很好的在目标领域中展开应用；在第六章中，主要对机器学习模型的训练过程与训练结果进行描述，计算出模型分类的精确率和召回率，并针对分类效果给出评估；最后两章对本课题研究成果进行汇总，总结研究过程中的对错得失，最后向对本课题研究中给予帮助的老师同学致以感谢。

# 2 相关概念介绍

## 2.1 文本分类

文本分类[16]是指在预定义好的分类体系下，依据文本的特征（例如属性及内容），将给定的文本数据与一个或者多个类别相关联的过程。文本分类方面的研究会涉及到文本内容理解和模式分类等若干个自然语言处理和模式识别方面的问题，一个文本分类系统不仅是一个自然语言处理系统，也是一个模式识别系统，系统的输入内容是待分类处理的文本，系统的输出是与该文本所关联的类别。

总体来说，在国内外对于文本分类方面的研究共经历的下面4个主要的发展阶段：

1、第一发展阶段（1958年-1964年）：这一阶段的主要贡献是对文本自动分类的可行性进行了研究。

2、第二发展阶段（1965年-1974年）：这一阶段的主要贡献是进行文本自动分类的实验研究。

3、第三发展阶段（1975年-1989年）：在这一阶段中，成型了一些文本自动分类方法， 学者们的研究重点转向对已有算法的深入探索与改进。

4、第四发展阶段（1990年至今）：这一阶段为面向全网的文本自动分类方面的研究。

## 2.2 情感分类

情感(Sentiment)，是指人们对于某件事物产生的内在的主观感受。情感分类（有时被称为倾向性分析）是指根据文本内容所表达出来的含义的不同，将不同的文本划分成包括两种（二分类）或多种类型（多元分类），是对文本发表人员观点或态度的划分。

情感分类问题不同于其他的一些分类问题，它既包含传统分类中的共性问题，同时它又有自己的独特性，例如，有些作者在进行情感信息表述时，他所表达出来的情感倾向可能会比较模糊，从而导致情感信息有着较大的多义性，也就是极性不明显的问题。针对这一问题，研究者们通过不懈的努力，在做了大量研究之后，提出了众多的情感分类方法，这些方法可以按照不同的归类方式划分:按所给文本中的情感特点分类，按机器学习方法划分。按所给文本中的情感特点的不同，情感分类研究可划分为数据不平衡性问题研究和领域相关性研究两类。若根据机器学习方法的不同，可以将文本情感分类分为有监督学习方法（训练数据既有特征(feature)又有标签(label)）、半监督学习方法（训练集同时包含有标记样本数据和未标记样本数据）和无监督学习（训练样本的标记信息未知）。

## 2.3 迁移学习

如今的许多机器学习方法要求训练数据和测试数据必须从同一特征空间（即本文中所说的域）中获得，并且需要具有相同或相近的分布。当分布情况发生较大的改变时，大多数的统计模型需要使用新收集的训练样本重新构建。但是在许多现实的应用中，重新收集所需要的训练数据来对模型进行重建，是需要花费很大代价的。在这种情况下，迁移学习(Transfer Learning)[18]的优势就显而易见了。首先，当我们发现待处理空间与当前用来构建模型的空间有着较大的差别时，我们不必丢弃现有工作而从新开始，使用迁移学习完全可以帮助我们复用当前的研究成果。其次，迁移学习有着强大的通用性和时效性，无论是在无监督学习、半监督学习还是有监督学习中，迁移学习都有着广泛的应用。如今，迁移学习可分为四种类型：基于样本的迁移学习、基于相关性的迁移学习、基于模型的迁移学习和基于特征的迁移学习。

2.3.1 基于样本的迁移学习

基于样本的迁移学习(Instance-based Transfer Learning)在使用中需要先对样本进行一种分类操作，将所有的样本按重要程度进行划分，用不同的权重值来代表不同的重要程度，另外还需要找到源域与目标域中共有的或相类似的数据，对这些数据放大之后，再到目标域中进行应用。这样的方法优点是方法简单，实现容易。缺点在于权重的选择与相似度的度量依赖经验，且源域与目标域的数据分布往往不同。

2.3.2 基于基于相关性的迁移学习

基于相关性的迁移学习(Relation-based Transfer Learning)适合应用在当两个与之间的关系比较相似的时候，例如以生物病毒的传播模型来预测计算机病毒的传播情况。当两个域相似的时候，它们之间会有共享某种相似关系，典型的方法是Mapping方法。

2.3.3 基于模型的迁移学习

基于模型的迁移学习(Model-based Transfer Learning)是一种比较直接的迁移学习方法，对于在源域中通过大量测试数据训练出来的模型，将它直接应用到目标域中，只需要对模型的参数进行微调，就可以达到一个比较好的训练效果，这样的方法的有点是可以充分利用两个领域模型之间存在的相似性，但参数调整过程较为繁琐，通常难以得到一个很好的参数组合。

2.3.4 基于特征的迁移学习

基于特征的迁移学习(Feature-based Transfer Learning)要求源域和目标域中含有一些共同的交叉特征，这时我们可以通过特征变换，将源域和目标域的特征变换到相同特征空间，使得该特征空间中源域数据与目标域数据具有相同的数据分布，然后在此基础上进行传统的机器学习。此类迁移学习方法的优点是对大多数领域适用，且效果较好，缺点在于难于求解，容易发生过适配的问题。需要注意的的是基于特征的迁移学习方法和基于实例的迁移学习方法的不同是基于特征的迁移学习需要进行特征变换来使得源域和目标域数据变换到同一特征空间，而基于实例的迁移学习只是从实际数据中进行选择来得到与目标域相似的部分数据，然后直接学习。

## 2.4 文本特征与特征选择

文本特征是文本中与文本表达的意图相匹配或能够代表文本意图的字、词、短语或其他元素，因此，在做自然语言处理或文本方面机器学习时，文本特征提取的准确与否将直接影响到模型的最终效果。这时就需要一个良好的特征选择算法，目前已有的特征提取方法比较多，常见的有：卡方统计(CHI)法、互信息(Mutual Information, MI)方法等，下面对这三种方法做简单介绍。

2.4.1 卡方统计(CHI)法

该方法使用一个卡方值来衡量某一特征t与某一类别c之间的关联度，卡方值的计算公式（2.1）如下：

其中，N表示训练集中的文本总量，A表示包含特征t且在类别c中出现的文档频率，B表示包含特征t且不在类别c中出现的文档频率，C表示不包含特征t且在类别c中出现的文档频率，D表示不包含特征t且不在类别c中出现的文档频率。对于计算结果卡方，其值越高，表示某特征项与该类别之间的关联度越高，以此来达到特征选取的效果。

2.4.2 互信息(Mutual Information, MI)方法

此方法的算法原理为：当一个特征与类型之间的关联程度越大时，其互信息值越大。互信息值的计算公式如下式（2.4）所示：

式（2.4）中的所有字母所代表的的含义与前面三个方法中使用的字母含义一致。从公式可知，当与无关时，与的值相等，最终互信息计算结果为0，此时特征与类型之间的关联程度最低。

每种文本特征提取算法都各有优势，在每一类问题中都应当选取最适合此问题的算法，在本课题中，将使用另外一种TF-IDF算法，这将在第四章中做详细介绍。

## 2.5 文本特征对齐

下面介绍在特征对齐过程中应用到的一些概念。

领域：对某一类别下事物的一种抽象的集合表示，在本文所讨论的此类文本分类任务中，领域即为两个行业中文本数据的集合。

源领域：有标签的文本所在的领域，在本课题中即电商领域。

目标领域：无标签的文本所在的领域，在本课题中即新闻领域。

枢纽特征：在源域与目标域中共同出现且出现频率较高的特征，这些特征可以作为在两个领域中做特征对齐时的桥梁，用于将两个领域中的相似枢纽特征联系起来。

非枢纽特征：除了枢纽特征之外的其他特征即为非枢纽特征。

相似枢纽特征：与枢纽特征相似度较高的非枢纽特征即为相似枢纽特征。

跨领域的文本特征对齐：通过一些算法，将两个领域中的非枢纽特征与枢纽特征进行关联，从而使在源领域中训练得到的文本分类器能够在目标领域中得以应用进而完成迁移学习，可以用如下四种算法来完成文本特征对齐任务：结构一致学习(SCL)算法，谱特征对齐 (SFA) 算法，特征集成及样本选择(SS-FA)算法[19]，Word2vec模型。

## 2.6 分类器

文本情感分类问题归根结底也是一个分类问题，分类问题的常用算法在文本情感分类中同样适用，所谓分类就是根据对象的特征将对象分配到相应的类别中去。目前的分类器或分类算法有很多，如支持向量机(Support Vector Machaines, SVM)分类器、朴素贝叶斯(Naïve Bayesian Classifier)分类器、决策树(Decision Tree)分类器、k近邻(k-nearest neighbor)算法等，下面对。朴素贝叶斯分类器、决策树分类器原理做简单介绍。

2.6.1 朴素贝叶斯(Naïve Bayesian Classifier)

朴素贝叶斯算法的核心思想是将特征和类别发生的联合概率作为给该文本分类的概率，因此要使用朴素贝叶斯算法首先需要对条件概率的相关内容有所了解，在条件概率中，在A事件发生的条件下B事件发生的概率记做P(B|A)，并且P(B|A)的计算公式如下式（2.5）所示：

在朴素贝叶斯算法中，假设训练集有d组数据，共有m个类别，其计算公式如下式（2.6）所示：

此公式表示，对于某一测试样例，分别算出此样例发生的条件下，属于类别的概率，去概率最大值对应的那个类别，即为朴素贝叶斯算法分类类别。

对于的计算，当x的各个属性之间相互独立时，若x有k个属性，可通过下式（2.7）计算得出

其中的计算方式如下：若是彼此无关的属性序列，对于训练集中的数据对象，属性下的属性值有n个，在类别下，相应的属性值有m个，那么的计算公式为（2.8）：

对于（2.15）式中的与，分别在训练集中统计所有样本中相应的事件发生的概率即可。经过以上计算方法，以概率高低来进行分类，这样的分类结果能够比较贴合真实情况，但是，经过以上算法流程的介绍，可以看出，此算法比较适应于多属性，多分类任务，对于本文所讨论的面向文本处理的二分类任务，作者认为并不适合。

2.6.2 决策树(Decision Tree)

决策树(Decision Tree)是一类常见的机器学习方法。顾名思义，决策树就一种基于树的结构来进行决策的算法，样本集中的全部数据从树的根节点开始判断，没经过一个节点进行一次判断，根据判断结果将该样例划分到相应的子树中去，重复上述过程，直至到达叶子节点，这时每一个叶子节点都有一个唯一对应的分类类别，所有叶子节点中的分类类别通常是存在重复的。决策树可以是二叉树也可以是多叉树，关键在于每一个属性存在几类属性值，以二分类任务为例，决策过程如图2.1所示。

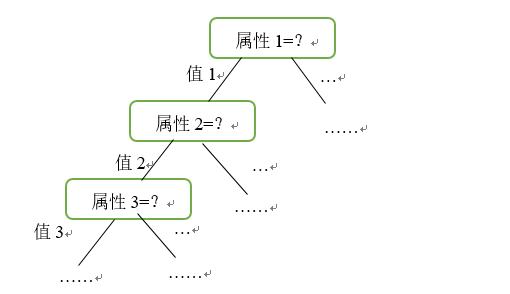


图2.1 二分类决策树示意图

一棵决策树的形成过程包括两个阶段：构造与修剪。构造就是选择何种属性作为节点，确定每个节点的划分，最终生成一颗完整的决策树的过程；修剪就是对决策树进行“瘦身”的过程，在这个过程中将一些多余的判断结果删除，使我们构造的决策树能够更加简洁，防止“过拟合”的情况发生。

“过拟合”是指模型的训练结果太好，以至于在实际应用的过程中，会存在“死板”的情况，导致分类错误，造成“过拟合”现象的原因是通常是因为训练集中样本数量较少，如果决策树选择的属性过多，构造出来的决策树一定能够极好地把训练集中的样本分类，但是这样就会把训练集中某一些数据的特点当成是所有数据的特点，这就使得这个决策树在真实的数据分类中会出现错误，也就是模型的“泛化能力”差。这里的“泛化能力”是指一个模型应对训练集中未存在的情况的能力，也就是“举一反三”的能力，当我们在训练模型时如果太依赖于训练集中的数据，那么就可能出现前述“过拟合”现象中存在的问题，也就导致了“泛化能力”差的现象。有“过拟合”自然就会存在“欠拟合”的情况，“欠拟合”是指模型的训练效果未达到最佳的情况，某些情况可能无法进行分类，这也会导致分类错误情况的发生。

## 2.7 本章小结

本章对本文研究方向中使用到的一些重要概念进行了详细介绍，首先对文本分类、情感分类、迁移学习此类的大型概念进行描述，其中对于一些在机器学习，文本分类任务中常用的算法，文本特征提取算法均进行了细致解读，以上均是作者本人在阅读大量文献之后，根据自身体会而总结出的算法机理。由于本课题任务是实现跨领域的文本特征对齐，因此在2.4节对于在文本特征提取中使用频率较高的四个算法进行了细致解读，因为研究的目的是实现跨领域的迁移学习，因此对机器学习中作者认为比较主流的分类器算法进行了阐述，以便为后面章节的研究工作打下基础。

# 3 文本特征对齐任务及数据预处理

针对在新领域样本中由于缺乏已标注数据而导致的分类效果差的问题，本章通过以“面向迁移学习的电商领域与新闻领域特征对齐方法的研究”课题，探讨在面向文本的迁移学习任务中，应实现的文本特征对齐任务、应有的工作流程、应获得的相关数据以及对应的数据预处理流程，为最终实现跨领域的迁移学习做好准备。

## 3.1 文本特征对齐任务

如今世界上存在各种各样的语言，不同语言之中存在许多的同义性和多样性，同一语言在不同领域中也会有着不同的特征，如前所述，要想使某一领域中的分类模型也能够在另一领域中使用，必不可少的一步就是实现跨领域间的特征对齐。

3.1.1 任务介绍

对于电商领域与新闻领域而言，本课题将对两个领域中用户评论的内容进行分析，以电商领域为源域，以新闻领域为目标域，选取两个领域中的对观点挖掘具有重要影响力的特定词性的词作为领域的特征，挖掘出两个领域中各自的枢纽特征与非枢纽特征，将出现在这些领域中的每一个相似枢纽特征对进行特征替换，从而进行两个领域中的特征对齐，并在特征对其后的源领域和目标领域数据上进行机器学习。

3.1.2 工作流程

要实现电商领域与新闻领域之间的迁移学习，首先要获得两个领域中的数据，这些数据应当满足如下要求：

1、选取数据的范围要广。要能够覆盖这个领域中的各个子类别，这样才能使这个数据集能够拥有该领域中的绝大部分特征。

2、数据要具有代表性。即一个领域中应当含有的词汇就应当大量出现，而这个领域中不应当出现的词汇的出现次数应该较少甚至为零。

3、数据量不宜过少。如果数据量过少，那么数据中包含的领域特征词会就过少，进而导致生成的模型在源领域中就不会有太好的效果，那么当迁移到目标域中之后，分类效果也不会很好。

4、文本长度适中，可以略长但不可以过短。文本长度长了会包含更多的领域特征词，这对于模型分类准确度的提升有帮助。但是如果文本长度过短，那么一个文本中所包含的特征词就会很少，这就很可能导致分类器判别出现失误。

在完成数据搜集之后，需要进行文本预处理工作，分为如下几个步骤：

1、文本净化处理。

2、文本分割。

3、标点符号处理。

4、停用词处理。

预处理完成后，即可开始特征提取任务，本文使用TF-IDF算法来完成特征提取任务。当源领域与目标领域中的特征词全部提取完成后，对于两个领域中的共有特征词，不做改动，对于两个领域中的非共有特征词，使用一种基于Word2vec工具的特征对齐算法将两个领域中的非枢纽特征进行对齐，原理是除枢纽特征外，对于目标域中的所有特征，使用Word2vec工具计算源域中与值相似度最大的特征词，当在迁移学习中调用分类器时，在目标领域中对文本中的特征词进行替换，使用替换完成后的文本带入分类器进行分类，本课题中选用支持向量机(svm)算法来构建分类器，将在源领域中训练得到的分类器应用在对齐之后的目标领域中，最后进行模型评估，评价本课题中算法的优劣。整体算法流程如下图3.1所示：

图3.1 本课题的处理流程示意图

3.1.3 评价指标

分类模型常用的评价指标有：精确率，召回率，准确率，错误率，F函数。其中，前两个指标在二分类任务中比较常用，后三个指标常用于多分类任务中，下面对他们分别介绍。

符号说明：TP(True Positive)指分类结果为正类的样例中分类正确的样例数，FP(False Positive)指分类结果为正类的样例中分类错误的样例数，TN(True Negative)指分类结果为负类的样例中分类正确的样例数，FN(False Negative)指分类结果为负类的样例中分类错误的样例数。

精确率：

召回率：

准确率：

错误率：

F1值是对精确率和召回率赋不同权值进行加权调和：

因为本课题所涉及的分类是一个二分类问题，因此使用前两个指标进行模型评估，精确率是指在预测为正类的样本中，预测正确的样本所占的比例，自然，精确率的值越高，分类器效果越好。召回率是指在真实的正类的样本中，预测正确的样本所占的比例，显然，召回率的值也应当是越高越好。

## 3.2 文本搜集

本课题名称为“面向迁移学习的电商领域与新闻领域特征对齐方法的研究”，自然需要搜集电商领域与新闻领域两个领域中的文本数据，结合3.1.2小节对文本搜集的要求，在两个领域中的文本搜集过程如下，数据采集过程均使用采集器完成。

3.2.1 源领域

目前国内市场上常见的购物平台有京东、淘宝、苏宁等，这些平台用户虽然很多，但是平台的商家中却存在一些不合理现象，有的店铺在与用户交易完成之后存在着向用户所要好评的现象，这种行为并不违规，但这种情况下的用户的评价内容的质量往往不会太高，有些对我们的研究甚至会出现负作用，这种现象在淘宝平台中最多，相比之下京东平台中对店铺的管理较为严格，本人认为选取京东的用户评论信息作为源域的数据集比较妥当。

为了满足前述要求中的范围广及代表性强的问题，京东主页上的商品分类有： 家用电器、手机、运营商、数码、电脑、办公、家居、家具、家装、厨具、男装、女装、童装、内衣、美妆、个护清洁、宠物、女鞋、箱包、钟表、珠宝、男鞋、运动、户外、汽车用品、母婴、玩具乐器、食品、酒类 、生鲜、特产、礼品鲜花、农资绿植、医药保健、图书、文娱、工业品、房产、汽车、艺术、教育、电子书、酒店、旅游、生活、理财、众筹、白条、安装、维修、清洗、二手、共52类，但是其中可以用于提取评论文本的类目只有前37类，因此从每一类中采集一部分商品评价文本，以此来保证所采集的数据的广泛性与代表性。

3.2.2 目标领域

新闻网站通常比较正式，国家把控比较严格，不存在源领域中的一些情况，本课题中选取新浪新闻中的用户评论信息作为目标领域数据集。研究发现，新浪新闻中的内容整体可以划分为两种类型，军事新闻与民生新闻，对于这两类，又可以分为国际新闻与国内新闻两个小类型，所以新闻领域数据集中的内容总共可以分为如下四类：国内军事、国内民生、国际军事、国计民生。

## 3.3 数据预处理

3.3.1 数据认知

源域与目标域数据均是一个两列的结构，第一列为文本内容(content)列，第二列为类别列(score)。因为在电商平台中，用户在发表完商品评论后会对该商品进行星级评分，范围在1至5之间，分数越高，该用户对该商品的满意度越好。在采集数据时，将用户的品论内容作为第一列，将用户的星级评价作为第二列。但是在新闻领域中，所有用户可以畅所欲言，没有打分的操作，因此只能采集到用户的评论信息，而无法采集到的评分值，这就需要采取人工标注的方式完成。

对于商品评论文本，长度过长的文本出现很少，但长度过短的文本出现很多，而且较长的文本在差评的评论中出现最多，这也是也一个在电商领域一个普遍存在的现象，这个现象并不会影响本课题效果的实现。

3.3.2 数据预处理

1、数据截取

本来采集到的用户数据很多，但是在后面章节的运行过程中发现，在文本处理这类问题的处理过程对电脑性能的要求实在过大，电脑在数以万计的数据量中无法结果，因此只选择2000条数据作为原始数据，具体提取过程如下。

新闻领域原始数据集的数据量共有203353条，累计有空值有8148条，其中总字数在20字到30字之间的评论文本数是30271条，从这30271个样本中选取2000条存入新的csv文件中，以这部分数据作为目标域的数据，同时，如前所述，这2000条新闻评价的评分需要人工添加的，然后再进行测试。另外，新浪网站的用户评论内容是经过utf-8编码之后的，还需要一个解码工作。

电商领域原始数据集的数据量共有68334条，属性列有两列，content列与score列，这两列中都存在空值现象，累计空值有11117个，其中总字数在20字到30字之间的含有评分的评论文本数是8332条，另外，因为此本课题是做二分类问题，而采集到了电商领域数据中，score值是以用户评分的星级为值的，因此需要进行转化，本课题中将评分为1或2的数据作为差评，将评分改为0，将评分为4或5的评分作为好评，将评分改为1，将评分为3的数据进行删除，操作完成后，累计6276条，同样从这6276个样本中选取2000条存入新的csv文件中。

2、文本分割

在文本分割任务中，使用jieba工具来实现分词任务，先对jiaba做一个简单的了解。

Jieba是目前常用的一个中文分词工具，jieba中文分词涉及到的算法包括：(1) 基于Trie树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图(DAG)。(2) 采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合。(3) 对于未登录词，采用了基于汉字成词能力的HMM模型，使用了Viterbi算法。

以句子“我来到北京清华大学”为例，Jieba支持三种分词模式：

(1) 精确模式：试图将句子最精确地切开，切分结果为“我/ 来到/ 北京/ 清华大学”，此模式适合文本分析。

(2) 全模式：把句子中所有的可以成词的词语都扫描出来，切分结果为“我/ 来到/ 北京/ 清华/ 清华大学/ 华大/ 大学”, 速度非常快，但是不能解决歧义问题。

(3) 搜索引擎模式：在精确模式的基础上，对长词再次切分，切分结果为“我/ 来到/ 北京/ 清华/ 华大/ 大学/ 清华大学”，提高召回率，适合用于搜索引擎分词。

此外，jieba分词还具有自定义词典的功能，用户可以将自定义的专有名词输入jieba中，这样当在分词时遇到这个词时，不会因为原本词库中没有而将该词汇拆分开，可以使用户的任务模型具有更高的正确率。

在对jieba有了一个大体上的了解了之后，可以对数据进行分词工作，将每个句子分完后，放入一个数组中，最后存入同一个文档中。经过jieba工具处理后，结果显示电商领域词汇共4025词，新闻领域词汇共6721词，信息统计的部分数据如下表3.1和3.2所示：

表3.1 分词后电商数据部分信息表

|  |  |
| --- | --- |
| 词汇 | 次数 |
| 我要 | 13 |
| 吐 | 6 |
| 槽 | 1 |
| 价保 | 18 |
| 空格 | 1961 |
| … | … |
| 蛮薄 | 1 |

表3.2 分词后新闻数据部分信息表

|  |  |
| --- | --- |
| 词汇 | 次数 |
| 不管怎样 | 3 |
| 私心 | 1 |
| ~ | 41 |
| 土耳其 | 60 |
| 在 | 606 |
| … | … |
| 葬身大海 | 1 |

3、停用词处理

Jieba的另一强大之处在于，他也能够进行停用词删除，但是jieba不自带停用词表，这个需要从网上自行查找。在将所有的停用词删除后，新闻领域词汇总数为6270词，电商领域词汇总数为3687词，重新对所有的词汇进行个数统计，统计出的部分结果如下表3.3和表3.4所示：

表3.3 去停用词后电商数据部分信息表

|  |  |
| --- | --- |
| 词汇 | 次数 |
| 挺好用 | 1 |
| 东西 | 96 |
| 质量 | 239 |
| 特别 | 57 |
| 期待 | 19 |
| … | … |
| 数额 | 1 |

表3.4 去停用词后新闻数据部分信息表

|  |  |
| --- | --- |
| 词汇 | 次数 |
| 中国 | 278 |
| 争霸 | 2 |
| 目的 | 12 |
| 全球战略 | 1 |
| 布局 | 1 |
| … | … |
| 葬身大海 | 1 |

观察数据可以发现一个现象，由于选取的数据量较少，去停用词之后的数据中存在大量的出现词数为1的词，那么这些仅出现一次的词也只能在一个句子中出现，这样的话，这些词对应的m、n、p、q的值也都相同，那么最终计算的TF-IDF值也可能相同，这时就需要设置一个TF-IDF值门槛来将这类干扰信息排除，这将在第4章中详细讨论。同时通过这两组数据发现另外一个有趣的现象，对于同样的字数（源领域与目标领域每个句子的字数都是在20-30字之间）、同样的数据量、同样的jieba分词工具和同样的停用词语表，新闻领域的分词结果总是比电商领域的多，这也从另一个维度上说明了两个领域的文本在特点上有着较大的不同，不过这并不在本课题的研究范围之内，将现阶段数据保存至磁盘，以便后续工作。

## 3.4 本章小结

本章主要进行了课题的任务梳理及数据预处理的工作。在第一小节中，对本本课题的完成步骤做了详细的梳理工作，对每一步的所要完成的任务、应选取的算法等方面做了详细介绍。在第二小节中，针对本课题所选择的两个领域，介绍了如何进行对文本的搜集能够满足广泛性、代表性、数量、长度的要求。在第三节中，先对采集到的数据进行了整体描述，随后着重介绍了数据的预处理过程，包括数据截取，文本分割，停用词删除，词汇统计四个方面。截止到本章，本课题的前期准备工作基本完成，为后期的特征提取，特征对齐，分类器构造等方面的工作做好了充分的准备。

# 4 特征提取

在前三章的基础上，本章进行文本特征提取工作，如前所述，文本特征是文本中与文本表达的意图相匹配或能够代表文本意图的字、词、短语或其他元素

，在第2章中已经对文本特征提取中的三个主流算法进行了讲解，本课题中使用了除了前述的三个算法之外的另一个主流算法——词频-逆文档频率(Term Frequency–Inverse Document Frequency, TF-IDF)统计法

## 4.1 算法选择

在阅读相关文献资料后，对TF-IDF算法及其优势有了较为深刻的了解。

词频-逆文档频率这个概念由两个词组成：词频(Term Frequency, TF)、逆文档频率(Inverse Document Frequency, IDF)。

词频是即一个词在文档中出现的次数，显然，一个词在文章中出现次数越多，那说明这个词肯定有着很大的作用，但是每篇文档中出现次数最多的词通常是形如“的”、“是”这类的词，我们称之为“停用词”，这样的词显然对我们的分析和统计没有什么帮助，反而有的时候会干扰我们的统计，因此我们就需要把这些没有用的词给去掉，可以利用现有的停用词语料库来完成这个任务。

获取到词频后，又会出现另一个问题：有些与本领域无关的词的词频与其它与本领域有关的词的词频一样，但这个词并不是特征词。解决这种问题的方法是引入一个权重，来衡量一个词是不是常见词。如果某个词比较少见，但是它在本领域的文本中多次出现，那么它就极其能够代表这个领域文本的特点。这也正是我们所查找的特征词。这个权重就是用逆文档频率(IDF)来表示，他的大小与一个词的常见次数成反比。设在样本数据集中词汇t出现的次数为m，样本数据集中的总词汇数为n，数据集中的语句总数为p，包含词汇t的语句总数为q，那么TF与IDF值的计算公式如下式（4.1）和（4.2）所示：

其中IDF的计算公式中，分母为的原因是防止分母为0的现象出现,底数值的选择不做强制要求，可计算要求灵活选择，为方便计算，本文中选择2作为底数。在计算得出TF与IDF的值之后，将二者相乘即为当前词汇对于整个数据集的TF-IDF值，如下式4.3所示.

某词汇的TF-IDF值越大，则说明这个词汇对于本领域的影响越大，就应当选取该词作为特征词。

与其他三个特征提取算法相比，TF-IDF算法有着明显的优势：算法简单，方便计算，比较符合实际需要，并且易于编程，因此对本课题条件比较符合，所以在本课题的实现过程中，本人选择TF-IDF算法来完成文本特征提取工作。

## 4.2 程序及流程说明

4.2.1 程序介绍

当前领域中，做文本特征提取的工具库有主要为sklearn，sklearn工具库既可以实现语句切分，又可以完成词频统计，也可以进行TF-IDF值的计算，许多学者也都比较喜欢使用这一工具库，但为了功能实现的方便性，也为了对TF-IDF算法能够有深入了解，在本课题中，自行设计了TF-IDF值的计算程序，每一个词汇与对应的总出现次数使用一个字典型变量wordsDict存放，数据集中的所有评论文本使用一个list型变量allwords存放。函数详情如下图4.1所示。

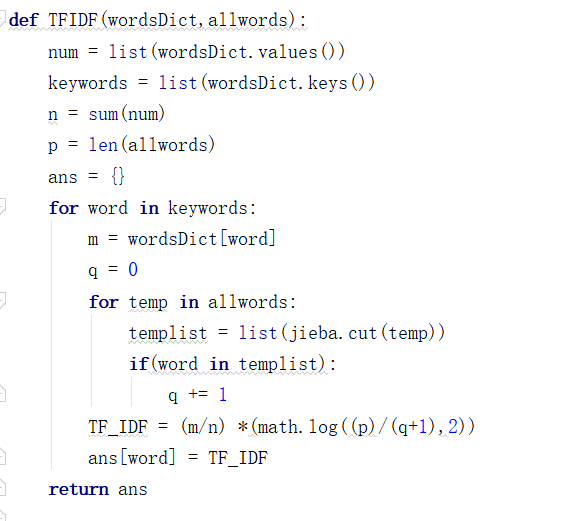


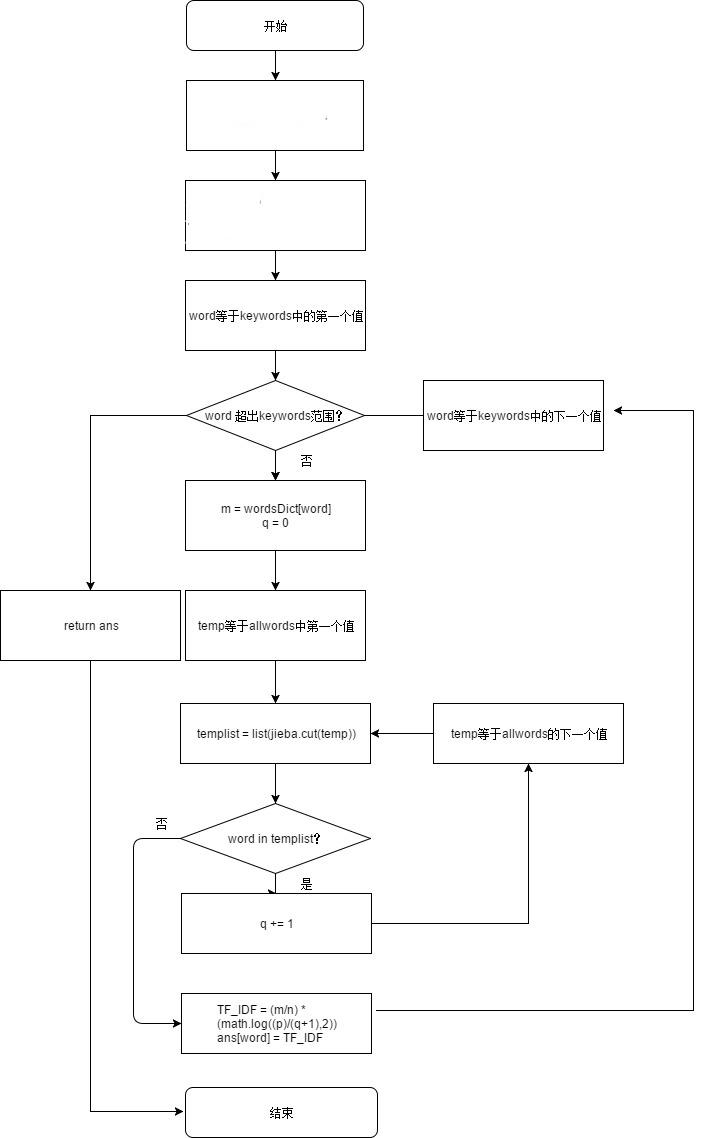
图4.1 TF-IDF值计算程序

4.2.2 程序流程说明

对于某个领域数据集，对所有的数据文本进行文本分割，去除停用词后，对每个词汇中包含的词汇的数目进行统计，统计结果存放在一个字典型数据变量wordsDict中，说有的文本数据则存放在一个list类型变量allwords中，将两个变量作为参数传入方法TFIDF中。

根据公式4.1-4.3中的变量知，在方法内应当求出词汇t出现的次数(m)、样本数据集中的总词汇数(n)、数据集中的语句总数(p)、包含词汇t的语句总数(q)，n与q的值很好计算，一个简单的求和操作即可完成，对于wordsDict中的每个元素，其对应的value值即为m的值，p的值则需要一个累加操作来完成，对于wordsDict中的每个key，需要再次遍历allwords中的每个值，再重新进行一次jieba分词操作，用来判断当前key值是否在allwords的当前句子中出现，若出现，则进行q的自增。

最后调用math函数，进行对数运算求得IDF值，再与TF值相乘，即可得到当前key的TF-IDF值，以键值对的形式存放到新的字典ans中，最后函数将ans返回，程序的运算流程如下图4.2所示。



**是**

**初始化num,keywords,**

**n,p,ans等变量**

**读入变量**

图4.1 TF-IDF程序流程图

4.2.1 程序性能分析

从程序结构看，程序中主要包含一个二层循环，所以时间复杂度应该是。看似并不大，但是其中包含了文本分割，字符查找等时间复杂度要求较高的操作，尤其是文本分割，一句话在分割完成之后词汇量会十多倍的增加，使得词汇量可以达到数以万计，双层循环下运算量将达到亿级，因此这个程序在实际运行中所消耗的时间还是相当大的。

## 4.3 程序运行情况

使用该算法分别对源领域与目标领域中的词汇的TF-IDF值进行计算，然后按照TF-IDF值由小到大进行排序。很容易发现，由于数据量较少，许多词汇的出现次数比较少，以1为例，这些仅出现一次的词也只能在一个句子中出现，进而导致，这些词对应的m、n、p、q的值也都相同，那么最终计算的TF-IDF值必然相同，这是影响最终模型分类精度的一个重要影响因素。但是，从另一种思路想，既然这类词的出现次数仅为1，那么这些词必然对该领域的研究没有贡献价值，即这些词必然不是特征词，因此可以放心的进行删除，排除在特征词的考虑之外，基于此分析，在对两个领域的特征词进行提取时，分别做了如下操作。

4.3.1 源领域

对于电商领域的数据，在去停用词之后的词汇总量为3687词，在程序的运行结果中，诸如“昨天晚上”、“配合”、“目标”、“男票”、“猛”这类的词汇的TF-IDF值全是0. 0006474094734163754，这显然就是上述提到的无关词语，比0. 0006474094734163754大的下一个TF-IDF值为0.0012188497908949384，这一等级的词汇多是诸如“省心”、“凑单”、“618”、“严实”、“飞快”此类的词汇，显然这些词已经具有了电商领域的特点，那么这些词就具有强领域相关性，因此，在对电商领域的特征词提取过程中，选择0. 001作为该领域的门槛，将TF-IDF值大于0.001的词汇留下，作为电商领域的特征词，结果显示最终留下的特征词只有1459个，非特征词有2228个。部分词汇的计算结果如下表4.1所示。

表4.1 电商领域特征词及TF-IDF值表

|  |  |
| --- | --- |
| 词语 | TF-IDF值 |
| 省心 | 0.0012188497908949384 |
| 很赞 | 0.0017474232254439314 |
| 秒杀 | 0.0018282746863424076 |
| 超值 | 0. 0022462799469415116 |
| … | … |
| 不错 | 0.06010613648758296 |

4.3.2 目标领域

对于新闻领域的数据，在去停用词之后的词汇总量为6270词，在程序的运行结果中，诸如“棋局”、“无休止”、“千真万确”、“繁荣”、“不用说”这类的词汇的TF-IDF值全是0.0006700723797267733，这明显也是无关词语，比0.0006700723797267733大的下一个TF-IDF值为0.0012615049126133025，这一等级的词汇多是诸如“争霸”、“自作自受”、“核导弹”、“朝鲜人”、“isis”此类的词汇，这些词就具有强领域相关性，因此，在对新闻领域的特征词提取过程中，同样选择0.001作为该领域的门槛，将TF-IDF值大于0.001的词汇留下，作为电商领域的特征词，结果显示最终留下的特征词只有1947个，非特征词有4323个。部分词汇的计算结果如下表4.2所示。

表4.2 新闻领域特征词及TF-IDF值表

|  |  |
| --- | --- |
| 词语 | TF-IDF值 |
| 争霸 | 0.0012615049126133025 |
| 强国 | 0.0023248608451456195 |
| 走狗 | 0.0028176730162243913 |
| 反击 | 0.0033812076194692695 |
| … | … |
| 支持 | 0.017302563941162877 |

## 4.4 本章小结

本章主要完成了TF-IDF算法程序的编写及运行工作。在第4.1小节中对TF-IDF算法进行了简要介绍，并说明选取原因，在4.2小节中对本章所用的程序进行了详细说明，最后在4.3节中给出运行结果，并且针对运行结果中存在许多相同且很低的TF-IDF值的现象给出了原因说明。截止本章，文本特征提取工作已完成，将所有实验数据进行保存，以便进行下一步的特征对齐工作。

# 5 特征对齐

为完成面向迁移学习的文本分类任务，在提取出源领域与目标领域文本特征之后，需要进行特征对齐的任务，进而完成迁移学习任务。

## 5.1 一种基于Word2vec的文本特征对齐算法

总所周知，特征对齐即通过算法，将源领域与目标领域中的特征词进行关联，从而使源领域中的分类器也能适应于目标领域的过程。针对文本特征对齐的任务要求，在当前领域内，许多学者均给出了自己的特征对齐算法，如John Blitzer[10]等提出的结构一致学习算法(Structure Correspondence Learning, SCL)、Sinno Jialin Pan[12]等提出的谱特征对齐 (Spectral Feature Alignment,SFA) 算法、Xia R等人提出的特征集成及样本选择(SS-FA)算法[19]等。在学习总结前人经验的基础上，本文中使用了一种基于Word2vec的文本特征对齐算法。

5.1.1 Word2vec简介

Word2vec 是在2013年由Google公司开发的一款用于计算词向量的工具。Word2vec可以用在拥有百万数量级的词典和上亿的数据集上进行高效地训练，Word2vec工具得到的训练结果我们称之为词向量(word embedding)，词向量可以很好地度量词与词之间的相似性。

在业内，有很多人会将Word2vec误认为是一个机器学习或深度学习算法，其实Word2vec是一个用于计算词向量的工具，Word2vec底层是一个浅层神经网络，其核心是两个基础模型：连续词袋(ContinuousBag-of-words, CBOW) 模型和Skip-gram模型，下面先简单介绍词袋(Bag-of-words, BOW)模型。

词袋(Bag-of-words, BOW)模型：

以文本t=“我爱中国”为例，对t进行分词处理，结果是“我/爱/中国”。当我们需要对t进行文本分析时，BOW模型会先将这三个词放入一个名为“词袋”的“容器”中，这个“容器”最开始是用set集合来表示，但是当需要处理的文本较多时，一些词汇会反复多次出现，这时使用set显然已经不合适了，所以现在模型中是使用数组来表示一个“词袋”。

BOW模型将文本t的分词结果存入数组words中，但是计算机实际上是无法直接对文本进行处理的，所以在将分词结果存入数组中的同时会建立一个索引，让每个词对应于一个编号，例如我——0，爱——1，中国——2。存放文本t的数组[“我”,”爱”,”中国”]就变成了[0,1,2]，这样对于文本t的处理将变得快速很多。词袋模型适用于一些离散、高维、稀疏的数据类型，但是词袋模型也存在一定的局限性，首先是维度问题，如果一个例子词典中包含10000个单词，那么每个文本需要用10000维的向量表示，当维度过多时，这对计算机性能的要求会相当高，其次因为词袋模型使用数组存放词汇的索引，因此无法保留词序信息，除此之外，它存在语义鸿沟的问题

连续词袋(ContinuousBag-of-words, CBOW) 模型：

一句话概括CBOW模型的功能就是通过文本的上下文或周围的单词来预测当前词，模型结构如下图5.1所示：

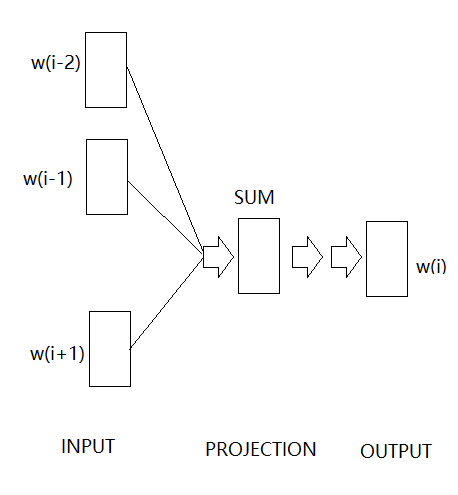


图5.1 CBOW模型结构示意图

CBOW模型的内部封装了一个神经网络模型，该模型是由前馈神经网络模型改进而来，该模型先将输入的上下文信息进行one-hot(独热编码)编码，将编码之后的内容作为该神经网络模型的输入，输入的内容通过一个权重矩阵W连接到隐藏层，隐藏层是一个N维向量，然后隐藏层又是通过另一个权重矩阵连接到输出层，最后输出的也是被one-hot编码的预测结果，

以前文t=“我爱中国”为例，若我们要预测“我”与“中国”两个词之间的词是什么时，我们的输入应该是2个词向量（“我”和“中国”这2个词的one-hot编码），输出则是数据集中的所有词汇的softmax概率，训练的目标就是找出softmax概率值最大的那个词，这个词就是待求的预测结果y。对应的CBOW神经网络模型输入层有2个神经元，输出层有词汇表大小个神经元，隐藏层的神经元个数可以自行指定。通过深度神经网络(DNN)的反向传播算法，我们可以求出DNN模型的参数，同时得到所有的词对应的词向量。这样当我们有新的需求，要求出某2个词对应的最可能的输出中心词时，我们可以通过一次DNN前向传播算法并通过softmax激活函数找到概率最大的词对应的神经元即可。至于CBOW其内层的神经网络模型的详细构造过程，在此不做额外介绍。

Skip-gram模型:

Skip-gram模型与CBOW模型是一个互为镜像的过程，其模型结构如下图5.2所示：

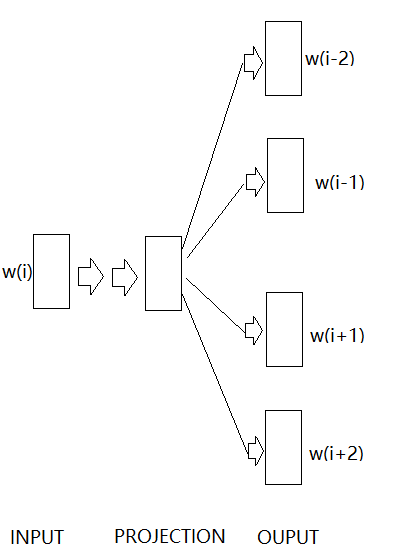


图5.1 Skip-gram模型结构示意图

同样地，Skip-gram模型的内层同样是一个神经网络模型，该模型的工作机制与前述的CBOW模型内层的神经网络模型工作机制也互为镜像，以同样的样例，模型的输入即为“爱”这一个词对应的one-hot编码，输出就是词汇数据集中softmax概率最靠前的2个词。

5.1.2 基于Word2vec的文本特征对齐算法

通过前述介绍已知，Word2vec有着能够查找某一词的最大相似词的作用，因此本文提出一种基于Word2vec的文本特征对齐算法，内容如下：

首先给出几个定义：

1、存放目标领域与源领域之间的特征对齐关系数据结构称为特征对齐库。

2、将源领域与目标领域中共同出现的特征词称为枢纽特征。

3、将所有特征词中没有在两个领域中共同出现的特征称为非枢纽特征。

对于枢纽特征，因为它们在源领域与目标领域中共同出现，本文认为枢纽特征对分类器性能的影响较小，因此这些枢纽特征不必进行特征对齐任务，因为即使对这些枢纽特征做了特征对齐，那么对其结果也只能是其本身，因为自己与自己的相似性一定是最大的，所以在特征对齐库中需要将这些枢纽特征进行剔除。

对于非枢纽特征，遍历目标领域中的所有非枢纽特征，使用Word2vec工具查找出源领域非枢纽特征中与之相似度最高的特征词，将这一对特征对齐信息存入特征对齐库中。

然后将目标领域中的所有文本数据拆分之后进行特征替换再拼接回去，拼接完成后的文本句子虽然人类读起来可能会非常奇怪，但是对于分类器是没有影响的，因为分类器是按照句子中的文本特征来分类的。

最后在源领域中训练分类器，迁移到目标领域后测试迁移效果即可。

## 5.2 算法实现

5.2.1 程序介绍

Word2vec在训练模型时需要输入全部的特征词，重复出现的特征词也要全部输入，所以在前面章节的基础上，程序需要将第四章求得的特征值信息读入，将两个领域的特征值合并在数组bothOfAllChWords中，还需要将所有的原始数据集读入，对数据集中的所有语句重新进行jieba分词，分别判断所有的分词结果是否在特征值数组bothOfAllChWords中已经存在，若已存在则将当前词语放入数组bothOfAllChWordsFF中，将以bothOfAllChWordsFF中的数据作为训练Word2vec模型时的输入。Word2vec模型的搭建程序如下图5.1所示：

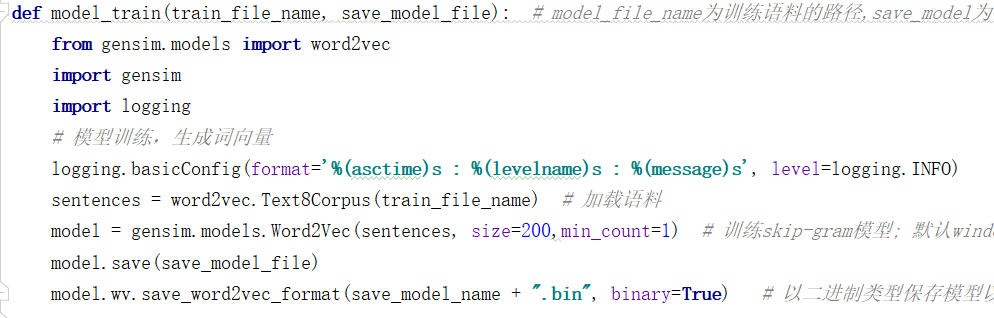


图5.1 Word2vec模型的搭建过程

Word2vec的模型可以算出bothOfAllChWordsFF中每个词与其他词之间的相似度，而且可以统计出与某一指定词相似度最高的若干特征词，并且同时还可以给出由高到低的排序结果。另外很容易可知，对于一个领域中的某一特征词，与此特征值相似度很高的其他词必定也是在相同的领域，而在第4.3.2节已知，在新闻领域最终所得到的特征词有1947个，因此若想求得在源领域中与目标领域中某一词语相似度最高的词，只需要将所求相似词语的个数设置为一个比1947更大的数字即可，在本课题中设置为2000，然后查找这2000个词中属于电商领域特征词集合中的第一个值，即为当前目标领域特征词的对齐结果，以后只需要以同样的方式遍历完其余所有的目标领域特征词即可。经过以上操作，最终，可以建立起两个领域特征词之间的之间的一对一关系，也就是5.1.2.节中所说的特征对齐库。建立特征对齐库的部分代码如下图5.2所示：

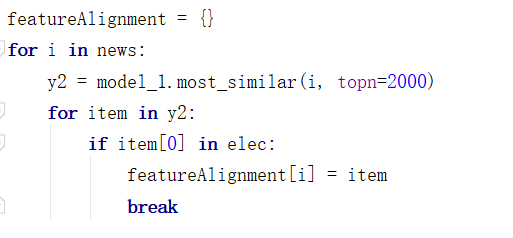


图5.2 特征对齐库的搭建过程

5.2.1 程序流程说明

整个特征对齐过程可用图5.3所示的流程图进行概括。

流

程

图

5.2.1 程序性能分析

从程序结构看，程序中主要包含一个二层循环，但是内层的循环次数最高为2000，并没有达到词汇总个数，所以时间复杂度应该是小于。虽然时间复杂度比第4章中的程序小一点，但是实际运行起来却是快了非常多，这也是Word2vec的一个强大的优势所在。

## 5.3 算法运行情况

在模型训练完成后，可以先使用目标领域数据查看一下训练效果，对于目标域中的所有特征词，查看全体特征词中与之相似度最高的三个词，部分结果如下表5.1所示。

表5.1 在全体特征词中对目标域特征词求相似特征词结果

|  |  |  |
| --- | --- | --- |
| 目标域特征词 | 相似特征词 | 相似度 |
| 争霸 | 学学 | 0.3138226270675659 |
| 朝鲜战争 | 0.25596365332603455 |
| 味道 | 0.25389885902404785 |
| 赢家 | 帝国 | 0.21515844762325287 |
| 想起 | 0.20881995558738708 |
| 应对 | 0.20824436843395233 |
| 自作自受 | 巡航 | 0.26888424158096313 |
| 坠 | 0.255778431892395 |
| 简单 | 0.24857649207115173 |
| … | … | … |

根据输出内容，首先可以发现，针对目标领域中的每个特征词在全部特征词集合中计算出来的前三个最相关特征词往往都是目标领域中的词，这很容易理解，因为一个领域中的词汇肯定首先与自己领域中的其他词的相关度最大，其次才能与其他领域的词有较大的相关度，这从一个方面证明了所得模型的正确性。另外，观察词与词之间的相关性，以人类的思维感觉所给的前三个词之间的相似性确实也较大，这也从另外一个方面证明了模型的正确性。

下面只需计算目标领域中特征词与源领域中的那个特征词的相似度最大，计算出的部分结果如下表5.2所示。

表5.2 在源领域特征词中对目标域特征词求相似特征词结果

|  |  |  |
| --- | --- | --- |
| 目标领域特征词 | 源领域特征词 | 相似度 |
| 争霸 | 味道 | 0.25389885902404785 |
| 赢家 | 没用 | 0.20473532378673553 |
| 挂 | 劣质 | 0.2697085738182068 |
| 自作自受 | 简单 | 0.24857649207115173 |
| 防 | 新 | 0.263325959444046 |
| 先发制人 | 浪费 | 0.2534712553024292 |
| … | … | … |

得出这个表，本课题的文本对齐任务已经完成，将输出结果保存至本地，以便后续使用。

## 5.4 本章小结

本章完成了在迁移学习过程中最重要的一步——特征对齐，通过这一步，已经使得源领域与目标领域中的文本特征趋于一致，进而使得跨领域间的迁移学习称为了可能。本章使用了Google公司的Word2vec工具实现了特征对齐，在第一小节中对Word2vec工具从原理上做了详细阐述，在第二节中编写了用于实现文本对齐工作的程序，在第三节中对程序运行情况给出说明，根据运行情况结合人工思考，初步可以确定对齐效果良好，这一章中的实现程序虽然看似简单，但是在实际编写过程中却是困难重重，导致这章内容的实现上花费了比较长的一段时间，不过也却是学到了很多。下一步将进行分类器模型的训练工作，最终结合迁移后的预测结果，对模型进行评测，这将在下一章节做详细介绍。

# 6 分类器构造与模型评估

支持向量机(SVM)分类器：目标是在训练集数据中寻找一个决策平面(decision surface)，这个超平面可以将训练集中的数据做最大程度上无误差的划分。支持向量机主要用于解决二元分类问题分。SVM的基本概念有如下三点：

线性分类：

二分类的问题通常使用一个线性函数（n是训练集样例的维数，R是实数集），对于函数的输出，我们可以规定：当函数f的输出值时，认为X的分类结果为负类，当函数f的输出值时，认为X的分类结果为正类。的通用形式如下式（2.7）所示。

f(x)即为所要寻找的超平面，这个平面可以将所有的输入（也就是x）分成正类与负类两半，如下图2.1所示。

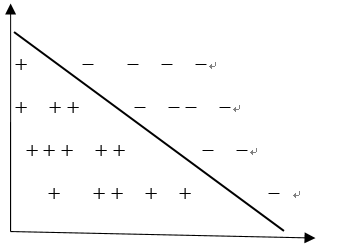


图2.1 f(x)所形成的超平面示意图

对于多分类问题，那么f(x)的值域应当是，将二分类中的线性模型向m类中转型就需要对每个类都关联一个相应的与，那么新的决策函数如下式（2.8）所示，这样可以为每一个类创建一个超平面。

线性不可分：

若当前问题是一个非线性问题，那么可以通过映射的方式，将样本集x映射到另一个空间中，然后在另一个空间中使用前述的线性分类，我们使用表示从空间X到空间F的映射，那么新的决策函数如下式（2.9）所示：

构造核(kernel)函数

核函数要求，对于, 满足

这里的是从空间X到空间F的映射。

在引入了核函数后，决策函数就可以用下式（2.11）表示，

此时，此算法的关键就在于核函数的建立，核函数要求适用的特征空间必须是对称的，也就是：

另外，核函数还必须满足Mercer定理：X的任意子集对应的矩阵必须是半正定的，综上是核函数的充要条件为：

经过以上介绍，可知，SVM算法最适用于二分类问题，且更适用于文本分类，目前文本分类方面的研究学者多数是使用的本算法来完成的文本情感分类的任务，因此在本课题中，也是用这种算法来完成迁移学习任务。

# 7 总结与展望

出现的问题

1 使用该算法分别对源领域与目标领域中的词汇的TF-IDF值进行计算，然后按照TF-IDF值由小到大进行排序。很容易发现，由于数据量较少，许多词汇的出现次数比较少，以1为例，这些仅出现一次的词也只能在一个句子中出现，进而导致，这些词对应的m、n、p、q的值也都相同，那么最终计算的TF-IDF值必然相同，这是影响最终模型分类精度的一个重要影响因素。但是，从另一种思路想，既然这类词的出现次数仅为1，那么这些词必然对该领域的研究没有贡献价值，即这些词必然不是特征词，因此可以放心的进行删除，排除在特征词的考虑之外，基于此分析，在对两个领域的特征词进行提取时，分别做了如下操作。（4.3）

2 停用词表中没有空格，导致空格的TF-IDF最大，好在两个都有，可以作为枢纽特征

3 部分字符超出解码范围的问题。

展望：

可进一步研究：

同时通过这两组数据发现另外一个有趣的现象，对于同样的字数（源领域与目标领域每个句子的字数都是在20-30字之间）、同样的数据量、同样的jieba分词工具和同样的停用词语表，新闻领域的分词结果总是比电商领域的多，这也从另一个维度上说明了两个领域的文本在特点上有着较大的不同（3.3.2）

# 删除了的

2.4.2 信息增益(Information Gain, IG)法

在了解信息增益法之前，需要先了解熵(Entropy)的概念，对于一给定的文本信息，熵指的是这一文本所包含信息量的多少。熵的计算表达式如公式（2.2）所示：

信息增益(IG)法就是根据某一特征在样本中所能提供熵值的多少来衡量此特征的重要程度，然后根据这个重要程度来决定此特征的去留。对某个特征来说，它的信息增益是指在样本中当不考虑此特征时的信息熵与考虑此特征值时的信息熵之间的差值。用公式表示如下：

其中，代表在样本中，类型的文本的出现概率，代表含有特征的文本的出现概率，代表不含有特征的文本的出现概率，代表某文本在包含特征的条件下又属于类型的条件概率，代表某文本在不包含特征的条件下又属于类型的条件概率，若用NUM来表示所有文本的类别数，则变量j的变化范围是。

2.6.3 k近邻(k-nearest neighbor)

顾名思义，k近邻是指k个距离最近的邻居，那么k近邻算法原理就是，对于一个待分类的样本，查找出距离这个待分类样本最近的k个已分类样本，在这k个已分类样本中，哪一个类别的样本拥有得最多，那么这个样本就会被分到该类别中去。如下图2.2所示，以正方形代表一个待分类的样本，以三角形和圆圈代表两中已分类的样本，以下图表示以个二分类任务，当k取3时，在距离正方形最近的3个样本中，三角形最多，所以该样本将被分派到三角形所在的类别。

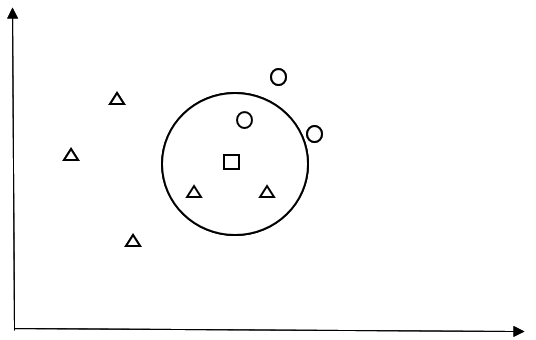


图2.2 当k=3时，knn算法的决策结果

如图2.3所示，当k值取5时，在距离正方形最近的5个样本中，圆圈最多，所以该样本将被分派到圆圈所在的类别。

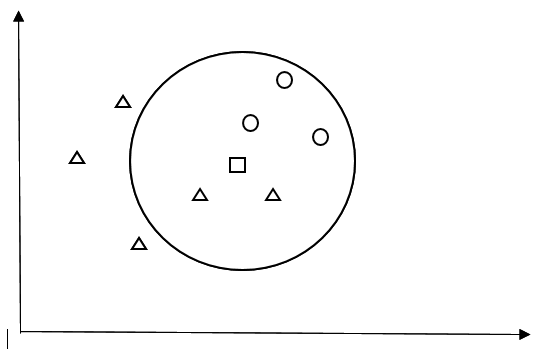


图2.3 当k=5时，knn算法的决策结果

经过两张图的对比可以发现，k值是一个影响决策结果的关键因素，k值的选择直接关系到分类器的运算结果，本算法另一个关键点在于点与点之间的距离怎么算，下面着重介绍这两点。

k值选取

k值的选取可以使用交叉验证法，将样本集按一定比例划分成训练集和测试集之后，先从一个比较小的k值开始，分别计算某一个k值下验证集合的方差，可以作出一个方差值随k值变化的折线图，方差值越小说明此时的k值的分类效果最佳，所绘制的折线图应当如下图2.4所示的形状。

图2.4 分类方差随k值变化示意图

此折线图中，方差值应当会随着k值的增大而先减小后增大，因为起初k值比较小，可用于判别的样本数量太少，所以分类误差会比较大，随着k值的增大，方差会逐渐减少，但到达某一k值之后，由于k越来越接近于样本集中的样本总数，分类效果会越来越差，错误率会再次提升，所以对于作出的折线图，我们应当取最低点对应的那个k值。

距离计算

度量空间中点之间距离的方式有多种，常见的有曼哈顿距离、欧氏距离等。

二者的计算公式如下：

曼哈顿距离：

欧氏距离：

在knn算法中，我们常用欧氏距离来计算两个样本之间的距离。

参考文献

1. 魏晓聪, 林鸿飞. 面向迁移学习的文本特征对齐算法[J]. 计算机工程, 2017, 43(2): 215-219, 226.
2. Turney P D, Littman M L. Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Transactions on Information Systems(TOIS), 2003, 21(4): 315-346.
3. Galavotti L, Sebastiani F, Simi M. Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization[C]// Research & Advanced Technology for Digital Libraries, European Conference, Ecdl, Lisbon, Portugal, September. Springer-Verlag, 2000.
4. Mladenic D, Brank J, Grobelnik M, et al. Feature selection using linear classifier weights: interaction with classification models[C]// SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004. ACM, 2004.
5. 林森, 唐发根. 基于Log似然比的特征选择算法[J]. 计算机工程, 2009(19): 62-64+67.
6. 杨凯峰, 张毅坤, 李燕. Feature Selection Method Based on Document Frequency%基于文档频率的特征选择方法[J]. 计算机工程, 2010, 036(017): 33-35, 38.
7. 路永和, 李焰锋. 改进TF-IDF算法的文本特征项权值计算方法[J]. 图书情报工作, 2013(03): 92-97.
8. 曹鲁慧, 邓玉香, 陈通等. 一种基于深度学习的中文文本特征提取与分类方法[J]. 山东科学, 2019, 32(06): 106-111.
9. 卢晨阳, 康雁, 杨成荣等. 基于语义结构的迁移学习文本特征对齐算法[J]. 计算机工程, 2019, 45(05): 116-121.
10. Blitzer J, Mcdonald R T, Pereira F. Domain Adaptation with Structural Correspondence Learning[C]// EMNLP 2007, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22-23 July 2006, Sydney, Australia. 2006.
11. 孟佳娜, 段晓东, 杨亮. 基于特征变换的跨领域产品评论倾向性分析[J]. 计算机工程, 2013, 39(10): 167-171.
12. PAN Jialin, NI Xiaochuan, SUN Jiantao, et al. Crossdomain sentiment classification via spectral feature alignment[C]// Proceedings of the 19th International Conference on World Wide Web. New York, USA: ACM Press, 2010:751-760.
13. Chung F R K. Spectral graph theory [M]. American Mathematical Soc, 1997.
14. Zhang Y, Xu X, Hu X. A Common Subspace Construction Method in Cross-Domain Sentiment Classification[C]// International Conference on Electronic Science & Automation Control. 2015.
15. M.Simeon, and R.Hilderman, “Categorical proportional difference: A feature selection method for text categorization,” The Australasian Data Mining Conference. pp. 201–208, 2008.
16. 宗成庆. 统计自然语言处理[M]. 北京: 清华大学出版社, 2013, 416-417.
17. 肖明. WWW科技信息资源自动标引的理论与实践研究[D]. 中国科学院文献情报中心, 2001, 45-48.
18. Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on knowledge and data engineering, 2010, 22(10): 1345-1359.
19. Xia R, Zong C, Hu X, et al. Feature Ensemble Plus Sample Selection: Domain Adaptation for Sentiment Classification[J]. IEEE Intelligent Systems, 2013, 28(3): 10-18.
20. Zhang S, Liu H, Yang L, et al. A Cross-Domain Sentiment Classification Method Based on Extraction of Key Sentiment Sentence[M]// Natural Language Processing and Chinese Computing. Springer-Verlag New York, Inc. 2015.
21. Zhou G, He T, Wu W, et al. Linking Heterogeneous Input Features with Pivots for Domain Adaptation[C]// International Conference on Artificial Intelligence. AAAI Press, 2015.
22. 李鼎宇, 胡学钢. 面向短文本的跨领域情感分类算法[J]. 小型微型计算机系统, 2018, 39(05): 1005-1009.