**大连海事大学**

**毕 业 论 文**

┊┊┊┊┊┊┊装┊┊┊┊┊┊┊订┊┊┊┊┊┊┊线┊┊┊┊┊┊┊

**二○一九年六月**

网络文本数据的爬取与分析

专业班级： 软件工程2班

姓 名： 王冶

指导教师： 李楠

信息科学技术学院

**摘 要**

本文针对互联网中的网络评论数据（以豆瓣影评为例），利用机器学习算法，分析其包含的个人情感，从而实现对于评论文本积极、消极情感的判断，以及灌水评论、垃圾评论的识别和剔除，达到舆情分析的初步效果。

第一，使用python对豆瓣网站的电影评论数据进行爬去，进行初步的数据处理并存入MonGoDB数据库中；第二，

提取数据库中的影评文本和对应的评论得分，为每一条评论标注标签；第三，使用jieba中文分词工具，对每一条文本数据进行分词处理、并生成TFIDF特征向量矩阵；第四，使用机器学习中的模型（朴素贝叶斯、支持向量机）以及深度学习中的卷积神经网络进行文本的情感分析。

本次实验基于Linux操作系统，以python作为开发语言，使用VsCode编辑器编写程序，借助于Sklearn机器学习工具包，以多种方法对文本的情感进行的训练分析，得到了理想的效果。

**关键词：网络爬虫；文本分类；情感分析；机器学习；深度学习**

**ABSTRACT**

Based on the network comment data in the Internet (taking douban film review as an example), this paper USES machine learning algorithm to analyze the personal emotions contained in it, so as to judge the positive and negative emotions of the comment text, as well as to identify and eliminate comments and garbage comments, and achieve the initial effect of public opinion analysis.

First, use python to crawl the movie review data of douban website, conduct preliminary data processing and store it in MonGoDB database. Second,

Extract the movie review text and corresponding comment score in the database, label each comment; Thirdly, using the Chinese word segmentation tool of jieba, word segmentation was performed on each piece of text data and the TFIDF eigenvector matrix was generated; Fourthly, the model in machine learning (naive bayes, support vector machine) and the convolutional neural network in deep learning are used for the emotional analysis of text.

This experiment is based on the Linux operating system. Python is used as the development language, and the VsCode editor is used to write the program. With the help of the Sklearn machine learning kit, the training and analysis of the text's emotion are carried out in various ways, and the ideal effect is obtained.

**Keywords:**

**Web crawlers; Text classification; Emotional analysis; Machine learning; Deep learning**

**目 录**

第1章 绪论 1

1.1 课题研究的背景及意义 1

[1.1.1 文本情感分析的研究现状 1](#_Toc307862865)

1.1.1.1基于情感词典的文本情感分析的研究现状 1

1.1.1.2 基于机器学习的文本情感分析的研究现状 1

1.1.1.2 基于深度学习的文本情感分析的研究现状 1

1.2 本章小结 ..1

第2章 基础知识及相关技术 4

2.1 视频图像预处理 4

[2.1.1 常用颜色模型 4](#_Toc307862870)

[第3章 视频图像预处理 7](#_Toc307862871)

[3.1 引言 7](#_Toc307862872)

[结论 10](#_Toc307862873)

[参 考 文 献 11](#_Toc307862874)

[致 谢 12](#_Toc307862875)

[附录1 1](#_Toc307862876)

**网络文本数据的爬取与分析**

**第1章 绪论**

**1.1 课题研究的背景及意义**

当今社会，在信息化的影响下，每个人都拥有着丰富多彩的网络生活，在享受网络生活的同时，网络也随之带来了许多问题。伴随着大流量app、网站的出现与流行，数以万计的网络数据进入人们视野，而垃圾评论、恶意灌水评论的大量存在混淆了人们的视听，错误引导舆论，甚至严重至歪曲事实真相。

基于这样的现实背景，分析网络文本评论数据、过滤垃圾评论、以及进行舆情把控便具有重大的意义。本文在这样的现实背景下，利用机器学习算法，对评论文本进行了特征提取和情感分析，从而初步实现对于网络文本数据的正负向情感分析、以及垃圾评论的过滤。

本文以提高文本情感分析性能为目标，通过研究机器学习算法（朴素贝叶斯、支持向量机） 、深度学习算法（卷积神经网络）并将其应用到中文文本分类这一问题，将有助于提高基于文本情感分析的网络舆情把控、用户评价等分析。因此，本文基于机器学习、深度学习的文本情感分析具有较高的科学研究意义和应用价值。

**1.2 文本情感分析的研究现状**

目前，文本情感分析主要有三类分类方法：

1. 基于情感词典的文本情感分析方法；
2. 基于机器学习的文本情感分析方法
3. 那么随着进年来深度学习技术的不断发展和深入研究，深度学习在自然语言处理这一领域也广泛应用开来。

在本节中，将简单介绍基于情感字典、机器学习、深度学习的文本情感分析的相关技术以及发展现状。

**1.2.1 基于情感词典的文本情感分析方法**

顾名思义，基于词典的文本情感分析方法通常根据人工搭建的情感词典，利用当前句子中存在的情感词、情感短语的情感加强、反转等规则来判断当前句子的情感类型。就研究现状而言，目前已经实现通过基于搜索引擎的方法、以及每个词语和情感词语之间的相关度等方法实现情感分类，并提高分类效果。不过基于情感词典的文本情感分析方法需要投入大量的人力，将非常耗费人力成本。

**1.2.2基于机器学习的文本情感分析方法**

机器学习算法主要分为两类，一种是有监督的机器学习算法，另一种是无监督的机器学习算法。

有无监督的区别在于：是否使用人工标注的数据作为数据集进行训练分析，近年来，包括使用朴素贝叶斯、SVM支持向量机、决策树等机器学习算法在文本情感分析问题上面都取得了非常不错的效果，朴素贝叶斯基于贝叶斯公式，将“已知特征求解类别”的问题转化成“已知类别求解特征”和“类别概率”的乘积问题；SVM基本思想是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。本文将使用这两种机器学习算法进行文本情感分析。

**1.2.3基于深度学习的文本情感分析方法**

还没有进行，暂时搁置

**1.3 论文组织结构**

本文的内容组织结构如下：

第一章：介绍当前网络生活中的大流量应用所带来的垃圾评论对人们生活的影响，以及对网络数据进行分析的意义。最后介绍网络文本情感分析主要的研究方法和现阶段的研究现状。

第二章：介绍本文所使用的技术、工具以及相关基础知识。包括使用python网络爬虫技术对豆瓣网站进行爬取，同时介绍本文爬虫的具体逻辑；除此之外，还将介绍如何过滤脏数据、冗余数据，以及正则表达式、BeautifulSoup等工具的使用；还将介绍自然语言处理中的文本预处理内容——分词、去除停用词；最后，本文将着重介绍所使用的机器学习、深度学习算法即朴素贝叶斯、支持向量机、卷积神经网络。

第三章、第四章、第五章未完继续。

**1.4 本章小结**

本章主要介绍了文本情感分析的研究背景、以及应用意义。同时也阐述了实现文本情感分析的三种主流方法的原理以及相应的应用现状，主要包括基于情感词典、基于机器学习、基于深度学习的文本情感分析方法。最后，本章给出了本文的内容组织结构。本文将主要使用基于机器学习、深度学习这两种方法实现网络文本的情感分析，同时还会涉及中文分词，生成词向量、文本预处理等相关内容。

**第2章 基础知识及相关技术**

网络文本数据的爬取与数据分析涉及到python网络爬虫、机器学习、深度学习、自然语言处理等相关内容，本文将对网络爬虫技术、自然语言处理中的中文分词，生成词向量、特征提取，机器学习中的朴素贝叶斯、支持向量机和深度学习中的卷积神经网络进行详细介绍。

**2.1 python网络爬虫技术**

网络爬虫是一个自动提取网页的程序，它为[搜索引擎](https://baike.baidu.com/item/%E6%90%9C%E7%B4%A2%E5%BC%95%E6%93%8E" \t "/home/wangye/文档\\x/_blank)从万维网上下载网页，是搜索引擎的重要组成。Python语言中提供了大量的库，能够帮助我们非常方便的从网络上爬取自几需要的数据。

**2.1.1 爬虫逻辑**

本文使用python-requests模块对豆瓣电影网站进行了爬取，共爬到电影评论数据4万条左右，数据内容主要包括电影的文本评论、对应评论的官方打分等相关内容。

本文需要情感丰富的文本评论数据作为数据集进行训练分析，而评论文本又不宜过长，故此选择了豆瓣电影的电影短评数据作为本次实验的训练集数据，网站数据如下图所示：



矩形框内便是需要爬去的电影短评数据，箭头指向的便是这条评论的得分数据，后期根据这个数据对标签进行人工标注。

具体的爬虫逻辑如下：



对于爬虫爬取下来的网络文本数据，要进行了特殊字符以及英文字符的处理，这里可以使用正则表达式、BeautifulSoup等工具，同时还要根据评论的得分情况，对其进行人工标注，1,2标注为0，表示为消极情感。4,5分标注为1，表示为正向情感。

**2.2 文本预处理技术**

**2.2.1去除标签、符号、表情**

由于爬虫爬取下来的网络数据包含大量的HTML标签、表情、符号、以及包括很多英文字符，这些字符的存在将大大降低情感分析的精准度以及效率。本文使用Python中自带的Re模块即正则表达式模块对英文字符、表情、符号进行剔除，同时使用BeautifulSoup去除HTML标签。

**2.2.2中文分词处理**

英文较中文来讲，由于空格的存在，天然不需要分词这一处理。而中文则不同。目前中文分词的工具包包括jieba、THULAC以及北大开源工具包pkuseg。目前分词工具实现的原理基本是基于规则、统计、语义、理解这四种方式实现。而本文使用jieba正是基于统计的分词方法，jieba分词工具包共有三种模式：精确模式、全模式、搜索引擎模式。而这三种模式具有不同的特点和优势，精确模式能够将句子精确地切分开，特别适合做文本分析。全模式对于句子词语扫描速度很快，但却无法解决歧义问题。搜索引擎模式，顾名思义更适合用于搜索引擎分词。本文的分词采取的是“精确模式”，这样利于文本的情感分析。

**2.2.3停用词处理**

停用词是一些不会影响当前句子感情的词或字，在做自然语言处理的时候，更多的采取将这些词过滤掉，从而提高后续工作的效率。目前对于停用词的处理，大都是基于停用词表，通过遍历的方式去除停用词。本文选用了哈工大停用词表，对豆瓣影评分词后的数据进行了停用词过滤。本文以5000条影评数据为例，对比了在未去除停用词和去除停用词之后的特征数，如下图所示：



可以看到再去除停用词后，特征数由26748下降到26517，当数据量达到几十万条乃至几十G的时候，去除停用词便显得尤为重要。

**2.2.4平凡词、独特词处理**

平凡词指的是在众多文本数据中均出现的词语，例如“电影”，由于出现的频数过高，其所代表的情感意义也就不是很突出，过于平凡。独特词指的是，在众多文本数据中某个词语仅出现很少的次数，也就是仅在少数文本数据中存在，那么它所代表也只是少数的特点，并不具备说服力。为了降低特征矩阵的特征数，进而提高中文情感分析的准确度和效率，本文采用统计词语在文件中出现的次数frequence、词语的文件占比率rate两项数据作为参考指标，设置阈值，将词语的文件占比率rate超过0.8，词语在文件中出现的次数frequence少于10次的词语全部剔除，极大的降低了特征数。如下图所示：



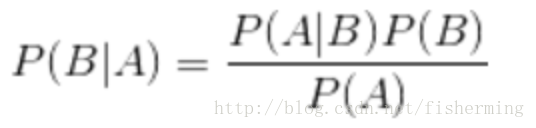
可见，在设置阈值，剔除平凡词和独特词之后，特征数成功下降到原来的十分之一。

**2.3 机器学习算法、深度学习算法**

**2.3.1朴素贝叶斯**

贝叶斯决策论是概率框架下实施决策的基本方法，对于分类任务来说，当所有相关概率在已知的理想情况下，贝叶斯决策理论将根据这些相关概率和误判损失进行最优分类，从而得出最优的分类类别标记。

而基于贝叶斯公式，我们知道：



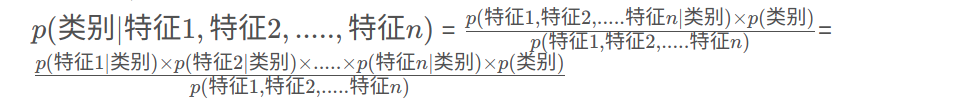
这其中，我们把P(B)成为“类别”的先验概率；而p(A|B)是样本A对于类别B的条件概率；P(A)称为用于归一化的证据因子，那么对于给定的样本A，p(A)与类别标记没有任何关系，那么问题：“在已知特征样本A的情况下，求解该样本为B的概率？”就将转换成如何根据训练数据估计先验概率P(B)、条件概率P(A|B)的乘积问题。

类先验概率P(B)表示的是各类样本占总样本的比例，反映了子样本的数量情况。根据大数定律可知，当训练集的样本数量充足时，各个样本满足独立同分布时，P(B)即可用各类样本出现的频率来表示。

对类条件概率P(A|B)而言，他涉及的是有关于所有特征样本A的联合概率，那么当各个特征相互独立的时候，类条件概率便可转换为各个子特征属性的条件概率乘积

那么在贝叶斯分类器的基础上，当我们假设各个特征属性相互独立时，那么每个特征属性都将对分类的结果产生影响，这时我们将贝叶斯分类器称为朴素贝叶斯分类器。“朴素”即代表特征之间独立假设的成立。

那么由上述条件可知，朴素贝叶斯的数学公式应用到特征分类领域为：



由于各个特征属性之间相互独立，那么条件概率便可以写成各个子特征的条件概率之积，而先验概率仍然保持不变。那么朴素贝叶斯分类器的训练过程便是，基于训练集D来估计类别的先验概率P(B)，并且估计每一个特征属性的条件概率。

朴素贝叶斯分类器共有三种模型，分别是多项式模型、伯努利模型、高斯模型。三个模型各有特点，也又所区别。其中多项式模型的特征为单词、特征值为该类单词出现的词频占百分比，并且在多项式朴素贝叶斯分类器中，特征向量多为离散型向量，应用于文本分类；伯努利模型以文本为特征，特征值为布尔型数据，标为0或者；高斯模型中，特征向量是连续性变量，并且假定所有特征的取值是符合高斯分布的。高斯模型适用于连续性变量预测。

**2.3.2 支持向量机（SVM）**

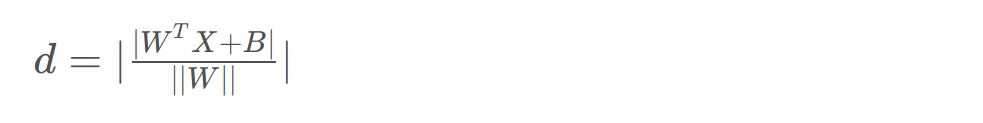
支持向量机算法模型在1995年被提出之后，得到了迅速发展，并在解决小样本、非线性和高维的模式识别问题中，均取得非常不错的效果。支持向量机根据其使用的核函数可分为：线性、多项式、高斯、拉普拉斯、Sigmoid类型的SVM。本文就情感分析问题上主要使用的是线性支持向量机，故此下文将详细的介绍线性支持向量机以及其数学推导过程。

支持向量机是一种有监督的学习算法。在二维平面上，散落着很多数据点，假设数据点仅有两类，那么我们可以找到一条直线对其进行分割，使不同类的数据点位于直线的两侧。同理在三维的空间中我们仍然可以找到一个面，将数据点分割开来，继而将维度上升至n维，那么也必定能够找到n-1的对象将n维中的数据分为不同的类别。这个n-1维的对象称为分隔超平面。在分割的过程中，离分隔超平面最近的点叫作支持向量。在实际应用中，人们通常希望找到最优的分隔超平面，所谓最优，就是指分隔超平面两侧的支持向量间的距离最大，当满足这个条件时，我们把它称为最大分类间隔超平面。

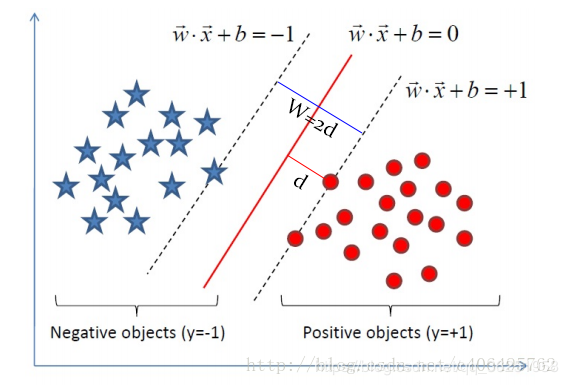
通过数学建模可知，在二维的情况下，分隔超平面的线性方程为：



那么支持向量到分隔超平面的距离为：



同时，如下图所示：



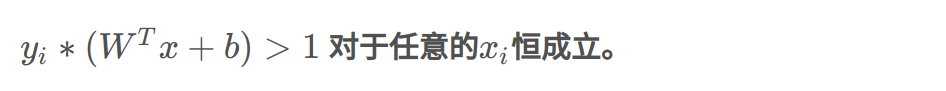
对于支持向量来说，其满足的线性方程为：



因此异类支持向量之间的距离为：



那么求解最大分类间隔的超平面即为求解两个异类支持向量之间的最大距离，也就是求解W的最小值。为了方便研究W的最值问题以及方便求导，将求解W的最小值转化成½|W2|的最小值，即MIN(½|W2|)。同时对于分隔超平面每侧的支持向量，均满足下面的关系式。



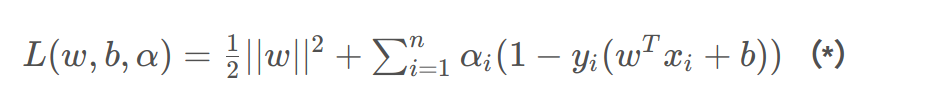
那么对于支持向量机的基本数学关系式为：



在约束条件：



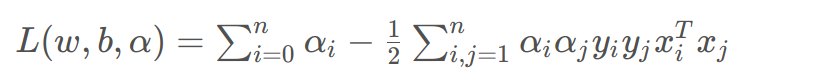
在后续的求解过程中，由于问题本身就是一个凸二次规划问题，故此可以使用拉格朗日对偶问题的求解思路进行解决。首先设置拉格朗日α因子，并规定α大于等于0。那么得到的拉格朗日函数为：



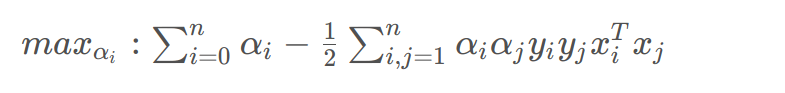
根据拉格朗日函数式，对W、α求偏导：



进而推导拉格朗日函数式为：



同时问题转换成：

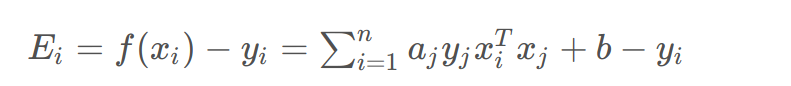


其应该满足的条件为：

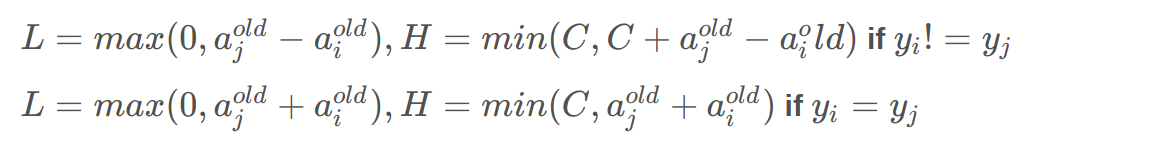


则问题由原来的拉格朗日函数式问题转换成求解合适α，使关系式取得最大值，那么求解α也就是smo算法问题，smo算法的数学推导如下：

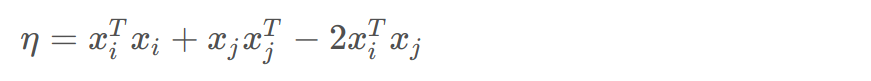
1. 计算误差：



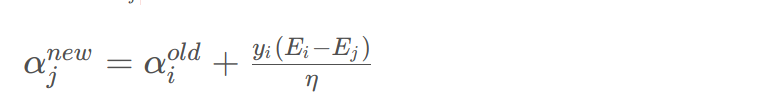
1. 计算上下界：



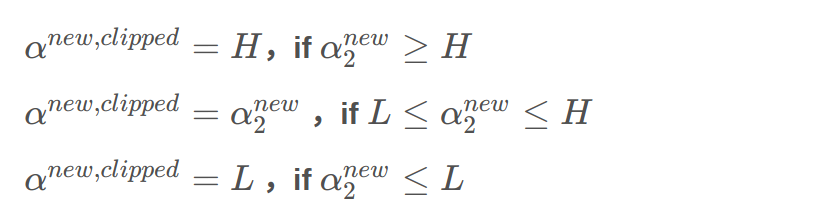
1. 计算η



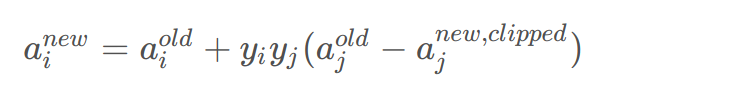
1. 更新αj



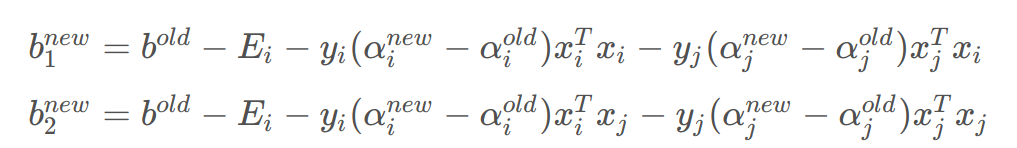
1. 修正αj



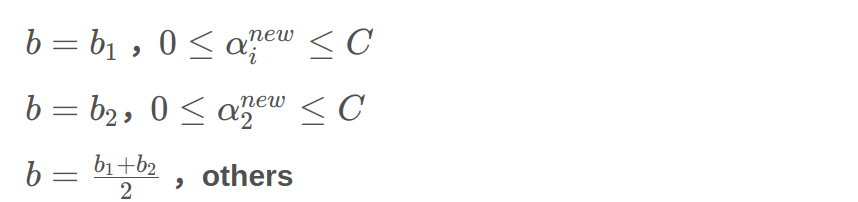
1. 更新αi



1. 更新b1和b2：



1. 根据b1、b2更新



SMO算法的工作原理是：每次循环中选择两个alpha进行优化处理。一旦找到了一对合适的alpha，那么就增大其中一个同时减小另一个。这里所谓的”合适”就是指两个alpha必须符合以下两个条件，条件之一就是两个alpha必须要在间隔边界之外，而且第二个条件则是这两个alpha还没有进进行过区间化处理或者不在边界上。

**2.3.3 卷积神经网络（CNN）**

卷积神经网络（CNN）在图像处理领域取得了很大的成绩，它的卷积和池化结构能很好提取图像的信息，而在 NLP 领域循环神经网络（RNN）则使用的更多，RNN 及其各种变种因拥有记忆功能使得它们更擅长处理上下文。但 NLP 领域很多方面使用 CNN 取得了出色的效果，比如语义分析、查询检索、文本分类等任务。这篇文章看看如何用 CNN 进行文本分类。

模型结构

模型结构可以通过下图一层层来看，总共由4部分组成，包括了输入层、卷积层、池化层和全连接层。

输入层

图中最左边的部分即为输入层，总的来说输入层就是句子对应的矩阵。一般不会使用 ont-hot 向量来表示单词，而是使用 k 维的分布式词向量。那么对于一个长度为 n 的句子，则构成一个 n × k 的矩阵。所以，可以设 xixi 为句子的第 i 个单词，它为 k 维向量。另外，根据对词向量的作用可以分为两种模式：静态和非静态。静态模式的意思就是我直接使用第三方发布的词向量或者自己训练的词向量来初始化矩阵，并且在每次训练的过程中不对反向误差传播产生作用，不改变词向量，整个训练过程词向量都是固定不变的。而非静态模式则不同，同样是使用词向量来初始化矩阵后，在此后的每次训练过程中，根据反向误差传播会对词向量进行微调，整个训练过程词向量都会更新。

卷积层

图中第二部分为卷积层，卷积层的作用就是用于提取句子的特征。主要是通过一个 h × k 的卷积核 w 在输入层从上到下进行滑动进行卷积操作，通过该卷积操作得到一个 feature map。feature map 的列为1，行为 (n-h+1)上图中输入层上红色框框就是卷积操作的卷积核，可以看到它是 2 × k 维的，运算后变为 feature map 的一个元素。除此之外，还可以将 h 定为3，此时卷积核变为 3 × k 维，如图中黄色框框部分。相同维度的可以有若干个参数不同的卷积核，所以最终在每种维度下都可以得到若干个 feature map。

卷积操作的意义是什么？可以看到它其实是根据 h 大小不同提取不同长度相邻单词的特征，这个其实可以跟 n-gram 语言模型对应起来。

池化层

图中第三部分为池化层，池化层的作用是对特征做进一步提取，将最重要的特征提取出来。这里使用的是 max-over-time pooling 操作，即取出 feature map 中的最大值作为最重要的特征，所以最终对于每个 feature map 池化后都得到一个一维向量，取最大值作为特征也解决了不同句子长短的问题，尽管短的句子会用 0 进行填充，但通过取最大值消除了该问题。

前面的通过卷积层的多个不同卷积核操作得到若干 feature map，而再经过池化层处理后得到若干个一维向量。

全连接层图中最后部分为全连接层，全连接层通过使用 softmax 分类器得到各个分类的概率。前面的池化层的输出以全连接的形式连到 softmax 层，softmax 层定义好分类。