编号 161710328

南京航空航天大学

本科毕业设计（论文）

题目 基于文本聚类的教学评估意见热点分析

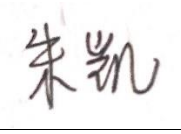
学生姓名	朱凯
学　　号	161710328
学　　院	计算机科学与技术学院
专　　业	计算机科学与技术
班　　级	1617103
指导教师	谢强 副教授

二〇二一年六月

南京航空航天大学

本科毕业设计（论文）诚信承诺书

本人郑重声明：所呈交的毕业设计（论文）是本人在导师的指导下独立进行研究所取得的成果。尽我所知，除了文中特别加以标注和致谢的内容外，本设计（论文）不包含任何其他个人或集体已经发表或撰写的成果作品。对本设计（论文）所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

作者签名： 

日　　期：　2021 年 5 月 19 日

南京航空航天大学

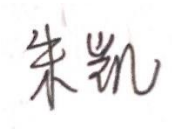
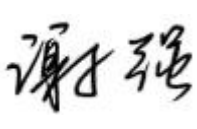
毕业设计（论文）使用授权声明

本人完全了解南京航空航天大学有关收集、保留和使用本人所送交的毕业设计（论文）的规定，即：本科生在校攻读学位期间毕业设计（论文）工作的知识产权单位属南京航空航天大学。学校有权保留并向国家有关部门或机构送交毕业设计（论文）的复印件和电子版，允许论文被查阅和借阅，可以公布论文的全部或部分内容，可以采用影印、缩印或扫描等复制手段保存、汇编论文。保密的论文在解密后适用本声明。

论文涉密情况：

√ 不保密

□ 保密，保密期（起讫日期：　　）

作者签名：  　　导师签名： 

日　　期：　2021 年 5 月 19 日　　日　　期：　2021 年 5 月 19 日

摘　要

教学评估作为高校进行教学管理的一个重要手段，在推动教学质量提升方面有着重要的作用，在教学评估中，学生对教师的评估一般包括打分和提交意见和建议。学生提交的意见和建议主要由任课老师自己自行查看，一般未进行综合的分析，所以这些意见和建议没能得到充分的利用。学校需要从这些大量的评估意见和建议中挖掘出有价值的热点问题，为教学管理和教学质量的持续改进提供支持。为此，论文采用了 K-means 的文本聚类方法对学校教学评估数据的意见和建议进行了热点问题的挖掘。

论文介绍了目前教学评估数据的意见和建议在高校教学评估中的重要性，分析了短文本聚类技术在教学评估数据应用中的意义。围绕这一课题，论文介绍了文本聚类领域的相关技术，对教学评估的意见和建议进行数据预处理，去除无用数据之后，建立了 TF-IDF 矩阵，设计了相关的 K-means 算法对教学评估数据进行热点问题的发现，进行了实验，并分析了实验结果。论文最后设计了系统的框架，进行了系统的功能设计和数据库表的设计，实现了数据清洗，热点问题发现等功能，并给出了系统的运行效果。

关键词：文本聚类，教学评估，文本挖掘，TF-IDF 矩阵，K-means 算法

ABSTRACT

As an important means of teaching management in Colleges and universities, teaching evaluation plays an important role in promoting the improvement of teaching quality. In teaching evaluation, students' evaluation of teachers generally includes scoring and submitting opinions and suggestions. The opinions and Suggestions Submitted by the students are mainly checked by the teachers themselves. Generally, there is no comprehensive analysis, so these opinions and suggestions can not be fully utilized. Schools need to dig out valuable hot issues from a large number of evaluation opinions and suggestions to provide support for the continuous improvement of teaching management and teaching quality. Therefore, this paper uses the K-means text clustering method to mine the opinions and suggestions of school teaching evaluation data.

This paper introduces the importance of opinions and suggestions of current teaching evaluation data in university teaching evaluation, and analyzes the significance of short text clustering technology in the application of teaching evaluation data. Around this topic, this paper introduces the related technologies in the field of text clustering, preprocesses the opinions and suggestions of teaching evaluation, after removing the useless data, establishes the TF-IDF matrix, designs the related k-means algorithm, finds the hot issues of teaching evaluation data, carries out experiments, and analyzes the experimental results. Finally, the paper designs the framework of the system, designs the function of the system and the database table, realizes the functions of data cleaning and hot issue discovery, and gives the running effect of the system.

KEY WORDS: Text clustering, teaching evaluation, text mining, TF-IDF matrix, K-means algorithm

目　录

第一章 绪论

1.1 背景和意义

教学评估已经成为中国各个高校评价教师教学质量的一个重要手段，教学评价目前大体分为客观打分和主观评价两种形式相结合进行。客观打分是指学生对老师上课的不同指标进行相应的打分，主观评价是指学生可以针对老师上课之中的一些问题提供一些具有建设性的意见，方便学校教务处发现其中的问题。客观评分很多时候由于主观的多种原因可能无法正确反映出问题所在，而学校大多时候可以从主观评价之中发现一些教学上的问题，具有参考价值。学校可以通过学生对于教师上课的评价从而得出一些学生们经常所反映出的热点问题，从而进行有针对性的解决问题，进一步提高教学质量。随着时间的推移，目前学校已经拥有数量庞大的教学评估数据，如何对这些数据进行正确的，完整的分析成为了目前学校所关注的问题。

首先，由于学校教务系统已进行教学评估很多年，已经拥有大量的教学评估数据，如何对这些数据进行正确的分析成为了一个很大的难题。其次，由于学生主观的因素，有些教学评估数据并不能客观的反映出问题的真实性。再者，教学评估数据之中还会有少量的无意义的数据，并没有什么建设性的参考价值。这些问题都会对准确的找出热点问题造成困难。所以我们必须找到有效合理的数据处理方法去解决问题。高效的清洗掉无意义的数据，准确的提取出教学之中所反映出的热点问题，找出海量数据之中有价值的信息。这是我们所关注的问题。

通过短文本聚类技术，我们能够从海量的数据中去除无意义的数据并且提取出在教学之中的学生们所关注的教学上的热点问题，把针对同一类问题的教学评估数据进行统一展现，可以让学校教务处快速高效地可以看见更具有建设性的意见，方便学校进一步的教学管理，可以针对所反映出的热点问题进行及时准确的教学调整，进行相应措施的教学改革，从而极大地提高教学质量。推动教学的更进一步改革。

1.2 研究现状

近年来，短文本聚类技术在很多行业之中已经有很多的实际应用了，微博，微信等的大型网络社交平台，凭借自身的优势，已经成为了人们生活之中常用的社交媒体平台，人们可以轻松的发表自己的评论，表达自己的想法，社交网络之中的评论巨量十分的庞大，模态多样等的大数据特点。这对于传统的文本检索与挖掘造成了很大程度上的困难。人们

可以在新闻、微博评论之中根据高频词寻找出热点话题。从而更进一步进行用户的数据分析。

目前，关于中文的分词以及文本的聚类已经有不少成熟的算法了，在文献[1]之中，介绍几种常见的文本聚类的方法并描述了这些聚类算法的具体步骤。关于文本聚类挖掘话题这一领域。目前也有不少学者使用过相关的算法，如在文献[2]，他们在论文之中对比分析了多个聚类算法之后，选择了 K-means 算法最终实现了微博话题的聚类，在文献[8]之中，采用了基于 TF-IDF 主题模型的词频矩阵，挖掘出了杭州市的社会热点问题。在文献[12]一文中，也提出了一种新基于重质心的词汇聚类的短文本相似度的新算法。使得短文本聚类算法技术有了更进一步的提升。

同时，也有学者将短文本聚类技术应用于教学评估之中，在文献[21]之中，他们将频繁词集共现网络用于教学评估数据之中，将教学评估数据进行分类，可以更规范化地看到学生们的建设性意见。在教学评估数据之中发现具有建设性意见，对于推动教学改革具有很大的价值。

国内也已经也有不会少发现热点问题的系统，比如将该技术应用于微博的热点话题发现或者是社会热点问题的分析，但关于用于教学评估的热点问题发现还是算是少数，本课题是要开发出一个可以应用于教学评估平台的可以挖掘出教学评估意见热点的系统，方便用户可以按照所选起始和结束学年与学期来更加直观的看到所选范围内教学之中的热点问题，并可以保存用户所查看到的数据的功能，方便学校进一步的教学管理，这就是本课题想要实现以及研究的内容。

1.3 本文主要工作

总体任务：学校每年产生大量教学评估数据， 对评估数据进行预处理，并通过文本聚类进行热点问题发掘，并对最终结果从用可视化的方式进行展现。并进行相应功能的系统开发实现。进一步反映教学情况，从而可以针对性地提高教学的质量。

本次毕业设计课题完成的工作：

（1）阅读相关算法的论文，学习并研究了已有文本聚类算法的原理。

（2）学习相应 Web 技术的开发，数据库管理软件的使用，以及相关编程语言的语法。

（3）对教学评估数据进行预处理并进行中文的分词。

（4）对分好词的数据进行特征提取并建立矩阵，计算每一个词在各个文本之中的 TF-IDF 词频矩阵。

（5）利用肘部算法对所拥有的特征矩阵训练模型选择出聚类数目

（6）利用 K-means 算法，结合所拥有的 TF-IDF 特征矩阵进行 K-means 聚类，并用图像的形式进行展现出每一类的前十个热点词汇。

（7）教学评估数据热点问题发现系统框架以及功能和界面的设计。

（8）完成代码的封装 WebAPI 接口并调用。

（9）教学评估数据热点问题发现系统的测试和运行，验证运行结果是否满足需求。

1.4 论文组织结构

本论文的组织结构如下：

第一章 阐明了课题当前的一些研究背景和意义，以及本文所用到的相关技术研究的主要现状，介绍了本文主要相关的工作，介绍了整体论文的架构。

第二章 介绍本文所涉及到的算法的原理，以及教学评估数据热点问题分析系统所需的相关技术。

第三章 介绍了教学评估热点问题发现算法的整体设计以及具体的时间步骤，并对所得的最终结果进行分析。

第四章 从系统需求的角度，设计了整体的框架，并对框架内各个模块的实现进行了简单的介绍。

第五章 对本课题之中所完成的工作进行总结，并对不完善的地方进行更进一步的设想和展望。

第二章 相关技术研究

本章主要介绍基于文本聚类的热点问题挖掘的一些相关理论和基础知识，一般来说通常分为文本数据的预处理阶段，文本的空间向量模型的建模，文本聚类三个主要的阶段。本章将围绕文本聚类过程之中所用到的相关技术进行介绍。

2.1 文本预处理

在真实场景之中的教学评估文本数据是无法直接用作热点分析的，因为在这些教学评估数据之中有少量数据是无意义的，并不能为我们进行热点分析工作做出什么参考性的价值。所以在聚类之前要对教学评估数据进行清洗，这可以大大提升我们的聚类效果，同时可以降低文本聚类的时间复杂度。我们一般可以先对文本数据进行去空和去重，再进一步去除常用中文停用词以及其他标点符号。其中，停用词是指那些没有什么实际含义的，所包含的信息量很少的词语。一般来说，停用词有两个类。第一类是在语法上面有一定的表现作用，但在语义上没有起到什么作用的词语。第二类就是一些含有特定功能的词语和一些表现力不是很强的词语。在进行文本聚类之前，在数据预处理阶段将停用词删除会提升文本聚类的效率。最常见的方法是构建停用词表的方法来去除停用词。

同时可以根据自身数据的特点再加入一下专门的停用词进行数据清洗，比如“老师”，“上课”，“讲课”等，更有助于本课题的教学评估热点发现。经历过简单的文本预处理阶段之后非常有助于我们接下来的文本聚类。

2.2 中文分词

2.2.1 基于统计的分词方法

中文之中的词语，是汉字和汉字之间的组合，所以在一篇文章之中，如果相邻汉字在同一时间出现的频率更高，它们就更有可能组成一个词。因此在分词中，我们需要计算词与词之间的互信息。互信息是指对文本中每个汉字的组合频率进行统计和计算，所以这种方法被称为基于统计的分词方法。互信息可以反映汉字之间的相关程度。当字与字之间的相关度高于一定水平时，我们认为这些字符可以组成一个词。这样子就省去了切分词语的工作，只用统计文本之中汉字组合的频度。

2.2.2 基于字符串匹配的分词方法

基于字符串匹配的分词方法是一个非常朴素的方法，这种方法只用一个“足够大”的词典之中的词语和想要去分词的文本进行比较。如果在词典之中找到了文本之中的某个子字符串，那么就识别出了一个词。从而达到了分词的效果。

2.2.3 基于理解的分词方法

基于理解的分词方法的本质是把计算机当作和人一样，模拟人的大脑来对语句进行理解和分词。他的主要思想是在分词的时候进行语法分析和语义分析，模拟人对于语句理解的成果。一般来说，这种方法有分词系统，语法语义系统和总体控制部分三部分组成。在总控部分的协调下，分词系统可以进行有关词语的语义信息来进行判断。但是，这种分词方法需要使用大量的语言知识和信息。所以目前还处于发展阶段。

2.3 文本的空间向量模型的建模

2.3.1 空间向量模型

向量空间模型（VSM：Vector Space Model）概念简单，这是一个非常经典的模型，在上世纪 70 年代提出。这个模型做到了从文字信息到数学模型的转换，可以完成很多大数据建模的工作。向量空间模型可以把文本内容转换为向量的运算，并且这样的模型还可以计算语义上面的相似度。通常将数据中所有文档出现的词映射为标签，而向量中每个维度代表该标签下所代表词的在这个文本之中的权重，完成了文本到向量之间的转换。

2.3.2 TF-IDF 权重

TF-IDF 算法通常用在空间向量模型上来统计每个词的权重，TF-IDF 是一种统计词在单个文档或者文档集中重要性的统计方法。本质上来说应该是一种加权算法。TF 词频（Term Frequency）是指该词在整篇文档中出现的频率，IDF 逆向文件频率,它的思想是如果一个词条 W 被更少的文档包含，那么 W 对区分文档内容的贡献度就越大，因此词条 W 的 IDF 值就越高。TF-IDF 为 TF 值和 IDF 这两者的乘积。公式为 TF*IDF。一个词 W 在一个文档 D 中的词频（TF 值）是指 W 在该文档 D 中出现次数（W;D）和文档 D 中总词数的比值：

$$TF(W;D) = \frac{\text{WordCount }(W;D)}{\text{DocSize }(D)} \tag{2-1}$$

1) 一个词 W 在整个文档数据集合中的逆向文档频率 IDF 值是指文档总数 N 与词 W 所出现文件数 Docs(W,D)做除法后得到的结果取对数：

$$IDF(W) = \log\left(\frac{N}{Docs(W;D)}\right) \tag{2-2}$$

2) 词 W 的 TF-IDF 值为：

$$\mathrm{TF}-\mathrm{IDF}=\mathrm{TF}(\mathrm{W};\mathrm{D})*\mathrm{IDF}(\mathrm{W}) \tag{2-3}$$

2.4 文本相似度计算

常见的聚类分析算法大多是根据元素之间的距离来执行聚类算法。例如常见的向量空间模型会按照一定的方法来将文本数据转化为空间向量，然后便可以使用几何学之中关于空间向量的一些方法和概念来定量的分析文本数据之间的相似性。计算两个文本向量d_i和d_j常见的方法通常有：

1）欧氏距离

欧氏距离也称之为欧几里得距离，这个公式可以真正意义上的计算出 m 维空间之中两个点之间的距离。公式如公式 2-4 所示。

$$\mathrm{Dist}\ (d_i,d_j)=\sqrt{\sum_{k=1}^{n}\quad f_k(d_i)-f_k(d_j)} \tag{2-4}$$

2）余弦相似度

余弦相似度是通过计算两个特征向量的夹角余弦值来表示两个文本之间的相似度。如果两个文档相似度越高的话，他们文本所表示出来的向量所计算出来的余弦值也就会越高。

$$\mathrm{Sim}\ (d_i,d_j)=\frac{\sum_{k=1}^{n}(f_k(d_i)*f_k(d_j))}{\sqrt{(\sum_{k=1}^{m}\quad f^2(d_i))^{+}(\sum_{k=1}^{n}\quad f^2(d_j))}} \tag{2-5}$$

2.5 常用文本聚类算法

2.5.1 K-means 聚类算法

K-means 聚类算法是一种经典的基于划分的聚类算法。它的主要思想是，将数据分为 K 组，并且随机选取 K 个对象作为聚类的中心点，然后计算每一个对象到所有 K 个中心点的距离，并把这个对象分配给距离它最近的那个中心点，并分配到该中心点所在的聚类之中。直到所有的样本都被分配到某一个聚类之中。这样就完成了聚类的工作。

K-means 算法的样例如图 2.1 所示。

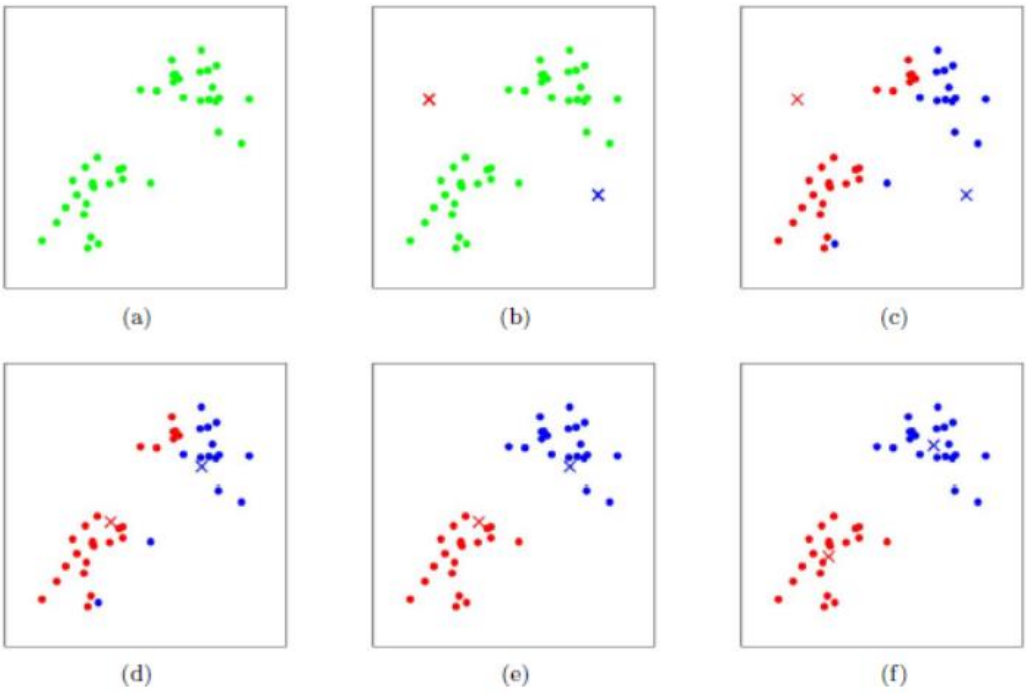


图 2.1 K-Means 算法结果图

通常情况下，K-Means 算法流程如图 2.5 所示。

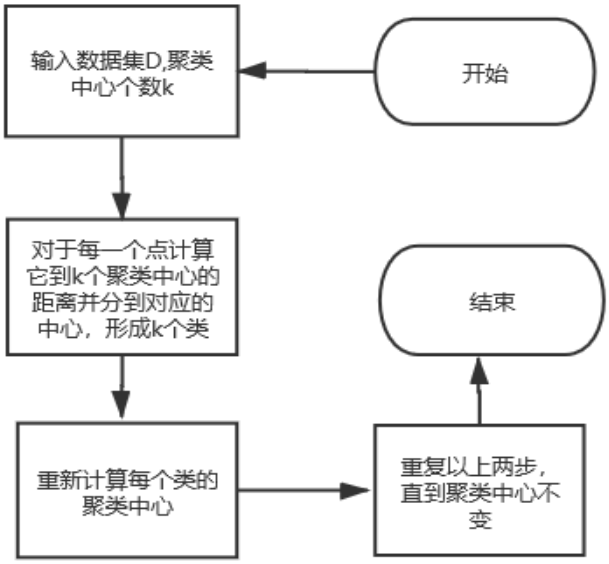


图 2.2 K-Means 算法流程图

2.5.2 DBSCAN 聚类算法

DBSCAN 聚类算法的本质是一种基于密度的聚类算法，这种算法是通过将密度在某一阈值之上的对象划分在同一个类别里面。这种算法的特点是，最终聚类的结果的形状可以是各种各样的，不仅仅是单一的按照距离区划分。与 K-means 比较起来，DBSCAN 聚类算法的聚类个数可以自己确定出来，并不需要提前设置。

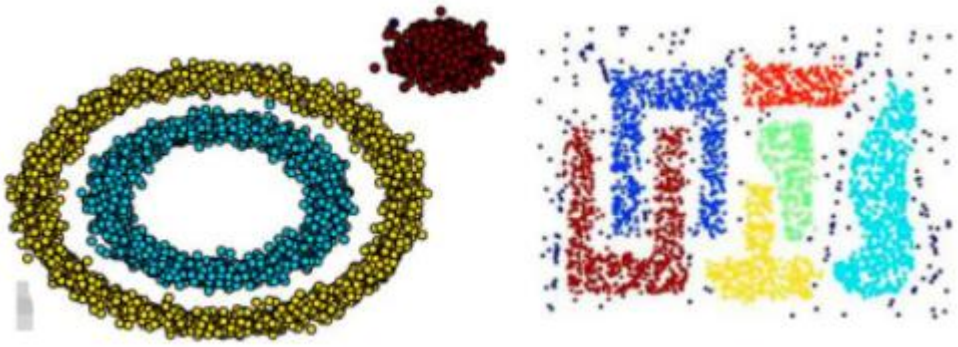


图 2.3 DBSCAN 算法结果图

一般来说，DBSCAN 算法流程如图 2.4 所示。

算法：DBSCAN，一种基于密度的聚类算法
输入：
D：一个包含 n 个对象的数据集
ε：半径参数
MinPts：领域密度阀值
输出：基于密度的簇的集合
方法：
1. 标记所有对象为 unvisited;
2. Do
3. 随机选择一个 unvisited 对象 p;
4. 标记 p 为 visited;
5. If p 的 ε – 领域至少有 MinPts 个对象
6. 创建一个新簇 C，并把 p 添加到 C;
7. 令 N 为 p 的 ε – 领域 中的对象集合
8. For N 中每个点 p'
9. If p' 是 unvisited;
10. 标记 p' 为 visited;
11. If p' 的 ε – 领域至少有 MinPts 个对象，把这些对象添加到 N;
12. 如果 p' 还不是任何簇的成员，把 p' 添加到 C;
13. End for;
14. 输出 C;
15. Else 标记 p 为噪声;
16. Until 没有标记为 unvisited 的对象;

图 2.4 DBSCAN 算法流程图

2.5.3 层次聚类算法

层次聚类方法对给定的数据集进行层次的分解，直到满足一定的条件。整体的流程可以看作是一颗树，按照聚类的方式的不同具体又可分为两种策略：

合并的层次聚类：初始阶段，将每一个对象都看作为一个聚类，然后将各个对象不断合并成为新的聚类。达到某个聚类的条件之后，聚类就会结束。

分裂的层次聚类：与凝聚的层次聚类相反，首先将所有对象看作是同一个聚类，然后再将这个大的聚类逐渐划分称为小的聚类，达到某个聚类条件之后，聚类就会结束。

层次聚类的算法示例图如图 2.5 所示。上方的流程为凝聚的层次聚类。下方的流程为分裂的层次聚类。

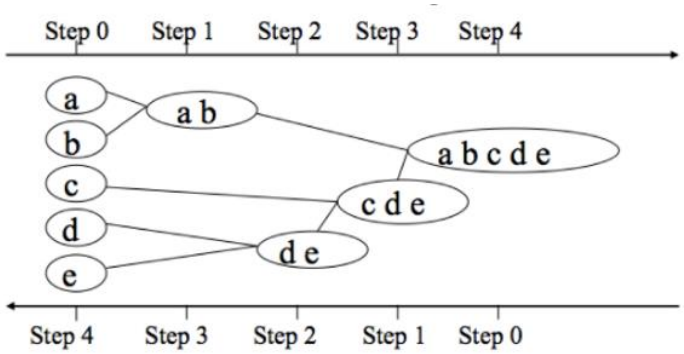


图 2.5 层次算法两种流程图

第三章 教学评估热点问题发现算法设计

3.1 整体流程

教学评估热点问题发现的分析需要经历大体上三个步骤，第一步为对教学评估数据的预处理，并且需要进行中文分词和去除停用词，第二步要对聚处理之后的结果进行建模，提取一些相应的特征，最后进行聚类算法并将最终的结果输出。教学评估热点问题发现算法流程如图 3.1 所示。

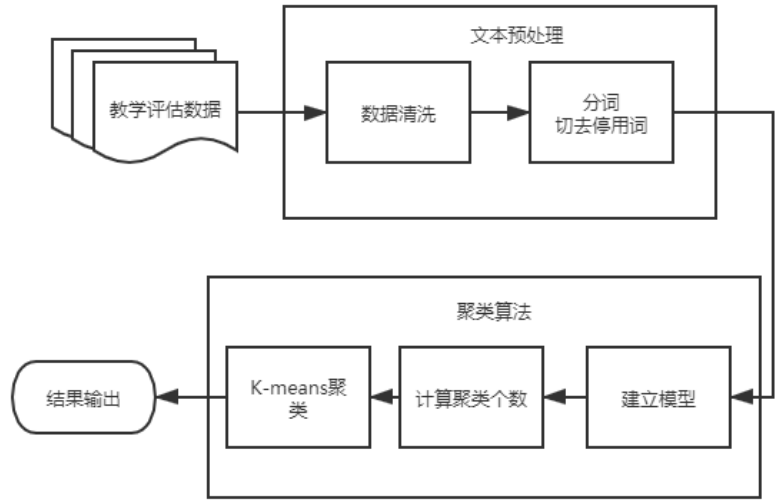


图 3.1 教学评估数据热点问题发现的任务流程

3.2 教学评估数据预处理

3.2.1 删除空数据

由于教学评估前期阶段，并未要求同学们都要强制填写教学评估意见，所以导致了少量的空数据的现象。所以我们的工作首先要对所有的教学评估数据进行去空，这样子会大大降低我们的时间复杂度，同时也会让我们的数据更加有意义。

3.2.2 删除重复数据

由于主观的因素，可能并不是所有的同学都会给出建设性的意见，所以也会出现少量的重复的教学评估数据，如图 3.2 所示。

思路清晰，结构严谨，内容丰富，幽默风趣，感觉挺好的！	2015-2016	1
思路清晰，结构严谨，内容丰富，幽默风趣，感觉挺好的！	2015-2016	1
思路清晰，结构严谨，内容丰富，幽默风趣，感觉挺好的！	2015-2016	1
思路清晰，结构严谨，内容丰富，幽默风趣，感觉挺好的！	2015-2016	1

图 3.2 教学评估之中重复数据样例

为了降低教学评估数据去重的时间复杂度，本课题采用了 simHash 算法。simHash 算法最大的特点就是将文本映射成为只有 01 的串，经过映射，只用对两个 01 串进行异或运算，计算为 1 的比特位的个数，这就是汉明距离的来源。一般来说，当所映射的两个 01 串之间的汉明距离越小，那就可以说明两个文本的相似度越高。就可以进行相应的去重。simHash 算法的流程如下。

（1）分词：将一条教学评估进行分词，对每个分词设置 1-5 的 5 个级别的权重。（这个权重可以是这个词出现的次数）。

（2）Hash：将各个分好的词映射成为 hasn 值，hash 值是由 01 组成的 n-bit 二进制串。

（3）加权：在得到了 hash 值的基础上，给所有特征向量进行加权，即 W＝Hash weight，且遇到 1 则 hash 值和权值正相乘，遇到 0 则 hash 值和权值负相乘。

（4）合并：将上述各个分词的加权结果进行计算和合并，变成只有一个序列串。

（5）降维：对于 n-bit 签名的累加结果，如果大于 0 则置 1，否则置 0，从而得到该语句的 simhash 值，最后我们便可以根据不同语句 simhash 的汉明距离来判断它们的相似度。

这样就完成了 SimHash 的去重算法。

3.2.3 建立停用词表

在教学评估的数据之中，也会产生少量含有停用词的无意义的数据，本课题也需要将这下数据进行清洗。

呵呵	2013-2014	2
呵呵	2013-2014	1
呵呵	2013-2014	1
呵呵	2013-2014	2
呵呵	2013-2014	1
呵呵	2013-2014	1
呵呵	2013-2014	1
呵呵	2013-2014	1
呵呵	2013-2014	2
呵呵	2013-2014	1
呵呵	2013-2014	2
呵呵	2013-2014	1

图 3.3 教学评估之中停用词数据样例 1

很棒	2013-2014	2
无	2014-2015	1
无	2014-2015	2
无	2014-2015	2
无	2014-2015	2
无	2014-2015	2
无	2014-2015	2
无	2014-2015	1
挺好的	2015-2016	1
挺好的	2015-2016	1
挺好的	2015-2016	1
挺好的	2015-2016	1
很棒	2013-2014	2
无	2014-2015	1
无	2014-2015	1

图 3.4 教学评估之中停用词数据样例 2

从网上下载已经建立好的各种版本的停用词表并最终整合，再根据聚类需要将类似于“老师”，“讲课”，“上课”等常用的词加入停用表以及子串表，更有助于我们的聚类以及看到更具有建设性意见的教学评论。图 3.5 为已提前建立好的停用词表的部分截图。

出
里
这门
没
不太
还行
行
一点
一门
稍微
棒棒
谢谢
棒
有时候
没什么

图 3.5 建立的停用词表

3.2.4 建立子串表

例如有教学评估数据之中会出类似于“6666666666666666666”或者“好好好好好好好好”这样子的数据，如图 3.6 和图 3.7 所示。这样子的数据不利于分词，其次通过删除停用词无法删除该数据。

好好好啊好好
好好好好啊好啊哈哦啊好
好好好好好啊好好好
好好好好好好
好好好好
好好好好好
好好好好好
好好好
好好好好
好好好

图 3.6 教学评估之中无意义的数据样例 1

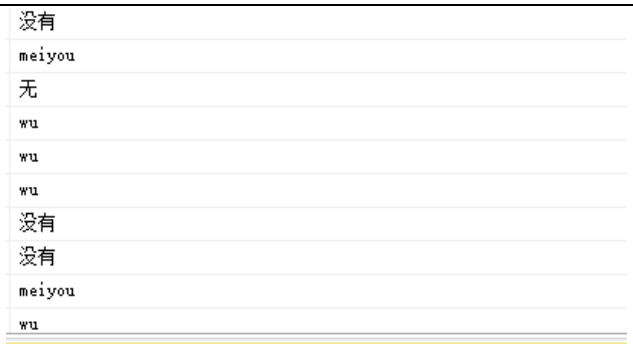


图 3.7 教学评估之中无意义的数据样例 2

所以建立一个无意义子串表，若包含有英文、数字或者特殊符号的无意义子串则删除该词。做到了更进一步的数据清洗。

3.2.5 中文分词

本课题之中采用的是基于字符串匹配的方法进行分词，将待处理的教学评估数据和建立好的中文词库进行匹配，按照最小匹配原则进行匹配识别词语。将预处理后的数据进行分词并用空格分开保存结果。如图 3.8 所示。

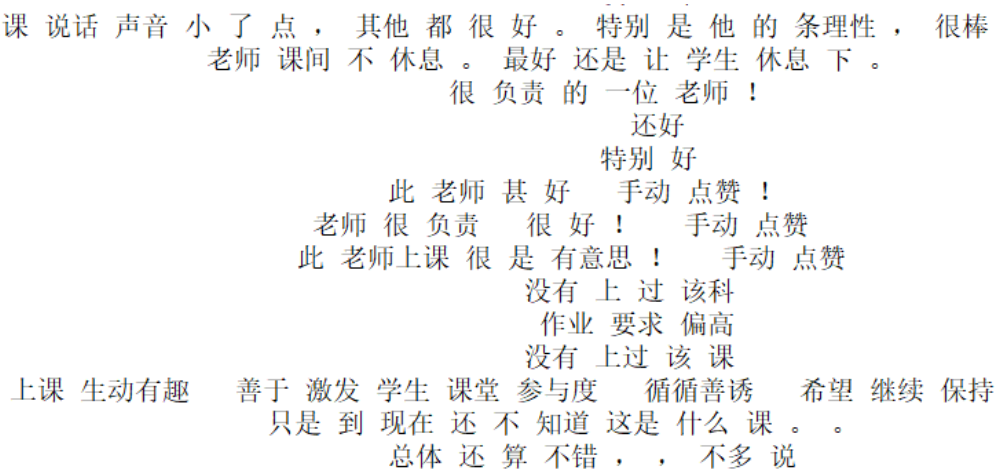


图 3.8 初步分词结果展示

3.2.6 删除停用词

中文分词之后，根据停用词表和无意义子串表，进行与教学评估数据的比对，删除停用词及无意义子串，经以上步骤处理后的结果如图 3.9 所示。

希望 加长 学时
希望 越 讲 越
讲
讲座 时间
周六 麻烦 阙 补课 辛苦
教材 太老 看不清楚
很亲
教学 态度 有条理 深入浅出 知识点
沙 教学 态度 严谨
课讲 人品
班任 大学 学术 严谨 教课 学生 关心 中 表率 班任 选择 学 直升机
教课 脾气 赞
博学 同学 提问 方式
清晰 明白 学生 极其负责 教材 写 位
教学 严谨 学生 关心
认真负责 教学实验 现场 课 相结合 教学方式 班主任
认真负责 告诉 查 文献 找 感觉 负责
教材 中 公式 推导 内容 跨度 较大 理解
书 太旧
李 幽默 风趣 收获
郭导 这课 关系

图 3.9 预处理后分词结果展示

3.3 教学评估数据聚类算法设计

首先，本科教学评估数据在经历了数据预处理阶段之后，就可以将所有的教学评估数据建立成 TF-IDF 矩阵模型,依据 TF-IDF 原理,对于各个不同维度除法的对应词进行处理，设X_{ij}为特征矩阵中第 i 行和第 j 列的值，实际意义为从该维度出发所对应第 j 列的所指代的词的 TF-IDF 值。

第二步就是要利用手肘法选择聚类中心数。手肘法的判断标准是 SSE 也就是误差平方和。其中误差平方和为：$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$ (3-1)

其中，C_i是第 i 个簇，p 是C_i中的样本点，m_i是C_i的质心(C_i中所有样本的均值)，SSE 是所有样本的聚类误差，代表了聚类效果的好坏。

手肘法的主要思想是:如果聚类数 k 越来越大,那么每个聚类的聚合程度会逐渐提高，从而误差平方和 SSE 自然会逐渐变小。当每个样本为一个聚类的时候，那么此时的误差平方和就是 0。并且，当 k 小于真实聚类数时，由于 k 的增大会大幅增加每个簇的聚合程度，故 SSE 的下降幅度会很大，而当 k 到达真实聚类数时，再增加 k 时它的回报就会变得非常的小，所以 SSE 的下降幅度会骤减，然后随着 k 值的继续增大而趋于平缓，也就是说 SSE 和 K 的关系图是一个手肘的形状，而这个肘部对应的 k 值就是数据的真实聚类数。所以，把这个方法叫做手肘法。

确定好聚类个数之后就可以对所有数据进行 K-means 聚类，利用 TF-IDF 矩阵作为模

型进行训练，随机选取聚类中心并最终得到聚类结果。

算法执行过程可分为三部分，分别为建立 TF-IDF 矩阵模型，选择聚类中心个数，K-means 聚类。下面为三个步骤的相似算法流程。

3.3.1 建立 TF-IDF 矩阵模型

输入：经过预处理的教学评估数据

输出：教学评估数据的 TF-IDF 矩阵模型

过程：

（1）选中第 i 个教学评论数据的第 j 个词，记为w_{ik}；

（2）计算该词的 TF-IDF 值，记为x_{ik} ；其中，i 表示为第 i 条教学评估。k 表示该词在所有词袋之中所处的维度。

其中：

$$TF(W;D) = \frac{\text{WordCount }(W;D)}{\text{DocSize }(D)} \tag{3-2}$$

$$IDF(W) = \log\left(\frac{N}{Docs(W;D)}\right) \tag{3-3}$$

$$x_{ik} = \text{TF(W;D)} * \text{IDF(W)} \tag{3-4}$$

（3）重复步骤（1）~（2），直到所有评论的所有词均计算完成；

（4）将所有计算的值变成矩阵；

算法流程图如下：

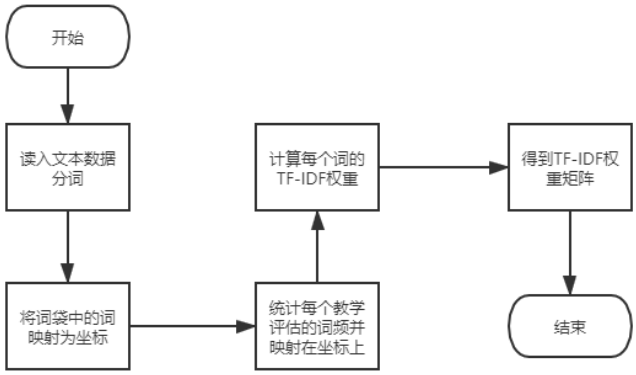


图 3.10 建立 TF-IDF 矩阵流程图

经过算法之后建立好的 TF-IDF 矩阵如图 3.11 所示：

```
[[0.00060914 0.         0.         ... 0.         0.         0.00136619]
 [0.         0.         0.         ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]
 ...
 [0.         0.         0.         ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]]
```

图 3.11 建立好的 TF-IDF 矩阵

3.3.2 手肘法选择聚类中心数

算法：手肘法计算误差平方和选取聚类中心个数

输入：待处理 TF-IDF 矩阵模型，待选择聚类个数 k

输出：选择不同聚类个数的 SSE 误差值

伪代码如下：

（1）For 聚类个数 k from 1 to n

（2）　　随机选取 k 个聚类中心

（2）　　For 聚类中的每个向量 i

（3）　　　　计算向量 i 到聚类中心的误差平方和

（4）　　　　加入聚类平方和之中

（5）　　End for

（6）End for

（7）选出肘点的聚类个数作为最终聚类个数

其中误差平方和为：$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$ (3-5)

随着聚类数量的增加，平方和会逐渐的变小。这个过程之中，会出现一个下降非常缓慢的点，也称之为“肘点”，此时的聚类个数可以作为我们真实所选取的聚类的个数。最终经过多次试验，当聚类个数为 20 个时，聚类效果最好。

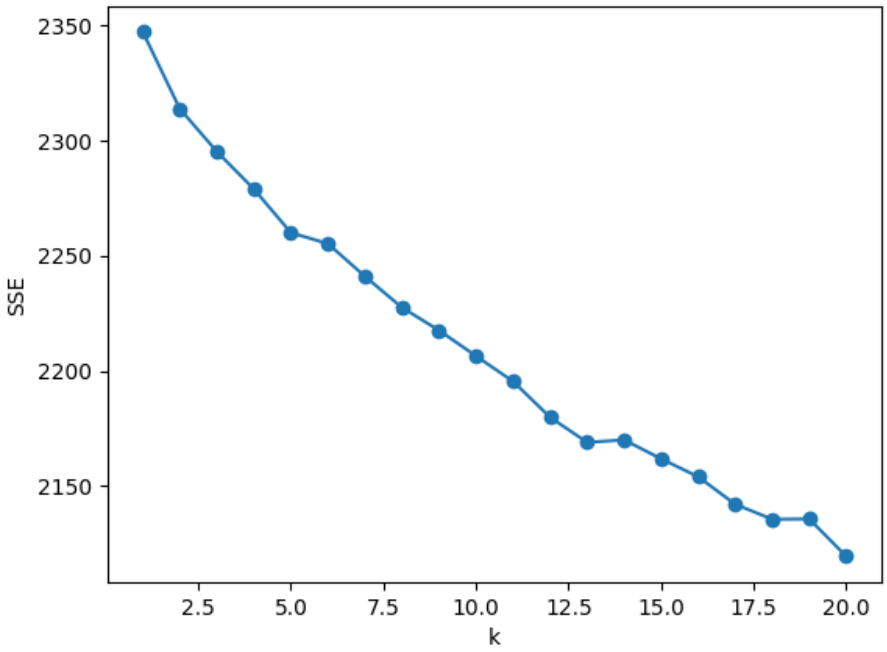


图 3.12 SSE 误差值和聚类个数坐标图

3.3.3 K-means 文本聚类

输入：待处理 TF-IDF 矩阵模型，所选择聚类个数 k。

输出：各个类别的评论以及词云图展示。

过程：

伪代码如下：

（1）Do 随机选取 k 个聚类中心

（2）For 样本中的每个特征向量 i

（3）　　For 每个聚类中心 j

（3）　　　　计算两个特征向量的余弦相似度

（3）　　End for

（3）　　将该向量分配给余弦相似度最大的聚类中心，并生成聚类标签

（3） End for

（4） 将聚类标签与教学评估建议和意见拼接

（5） 生成各个类别的热点词云图

其中余弦相似度为：$\mathrm{Sim}(d_i, d_j) = \dfrac{\sum_{k=1}^{n}\left(f_k(d_i) * f_k(d_j)\right)}{\sqrt{\left(\sum_{k=1}^{m} f^2(d_i)\right)^{+}\left(\sum_{k=1}^{n} f^2(d_j)\right)}}$　　(3-6)

算法流程如图 3.13 所示。

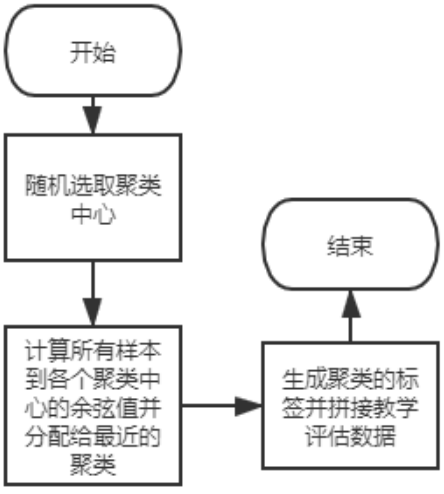


图 3.13 SSE 误差值和聚类个数坐标图

3.4 算法的实验及结果分析

3.4.1 实验的系统环境

客户端：Windows 10 系统电脑。

处理器：Intel(R) Core(TM) i5-7300HQ CPU @ 2.50GHz 2.50 GHz。

3.4.2 实验数据准备

选取南京航空航天大学 2017-2018 学年 2018-2019 学年，2019-2020 学年三年六个学期的评教数据进行实验，数据集包含 2672705 条教学评估数据，实验对每个学期进行数据处理分析，发现热点问题，并进行词频结果展示和词云图的展示。

3.4.3 中间结果的展示

（1）词频统计

先统计了分词之后的词袋之中所有的词频统计，并从高到低进行了排序，如图 3.14 所示。

```
耐心 -- 391 times
幽默 -- 407 times
互动 -- 425 times
负责 -- 441 times
教材 -- 463 times
认真负责 -- 481 times
有趣 -- 490 times
教学 -- 532 times
同学 -- 552 times
知识 -- 578 times
建议 -- 589 times
课程 -- 629 times
课堂 -- 666 times
讲解 -- 692 times
内容 -- 833 times
喜欢 -- 887 times
课 -- 1037 times
讲 -- 1454 times
学生 -- 1524 times
希望 -- 1990 times
```

图 3.14 词频展示

从这之中也已经可以看出一些热点词汇的展现。其实从热点词频之中可以初步的看出一些热点词汇。

进一步计算 TF-IDF 权之后的统计的结果如图 3.15 所示。

```
耐心 -- 0.09999941131203177 times
幽默 -- 0.10409145883375173 times
互动 -- 0.1086950122956867 times
负责 -- 0.11278705981740667 times
教材 -- 0.11841362515977163 times
认真负责 -- 0.1230171786217066 times
有趣 -- 0.12531895535267407 times
教学 -- 0.136060580097189 times
同学 -- 0.14117563949933895 times
知识 -- 0.14782521672213392 times
建议 -- 0.1506384993933164 times
课程 -- 0.16086861819761633 times
课堂 -- 0.17033147809159374 times
讲解 -- 0.1769810553143887 times
内容 -- 0.21304222409954593 times
喜欢 -- 0.22685288448535085 times
学生 -- 0.3897675264438271 times
希望 -- 0.5089484105139213 times
```

图 3.15 TF-IDF 权重展示

从这之中也已经可以看出一些热点词汇的展现。

3.4.4 数据词云结果展示

本节展示了三年六个学期的热点问题词汇词云图，并对每一学期的热点词汇进行问题分析。

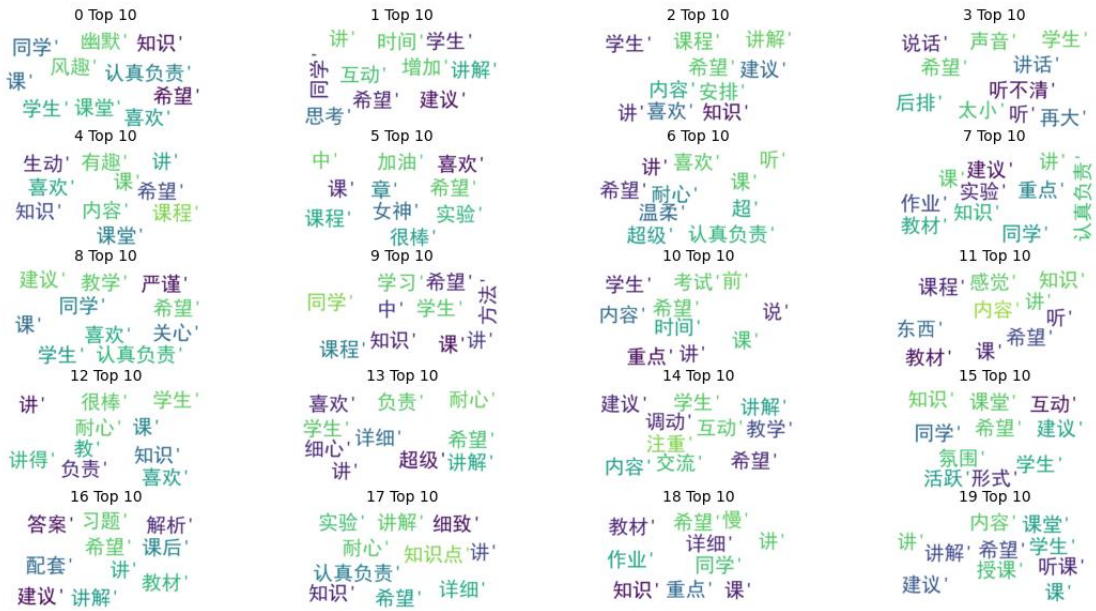


图 3.16 2017-2018 学年第一学期热点问题词云图

从图 3.16 可以看出在 2017-2018 学年第一学期，第 0 类热点问题之中，有“幽默”，“风趣”，“认真负责”，看得出同学们希望讲课风格幽默的老师，同时希望老师也要认真负责。第 1 类热点问题之中，有“时间”，“增加”，“互动”，“讲解”看得出同学们希望讲课

时需要多多增加互动，并且也要增加讲解的时间，可以帮助发现课堂上需要注意的问题。在第 17 类可以看到“细致”,“讲解”,“知识点”,“耐心”，所以分析同学们希望上课之中，老师会应该多多注重重点知识的讲解。在第 16 类热点问题之中，有“习题”,“解析”“讲解”“课后”等的词，那么就可以想到同学们希望可以有配套的课后习题教材，同时也要多一些习题上面的讲解。

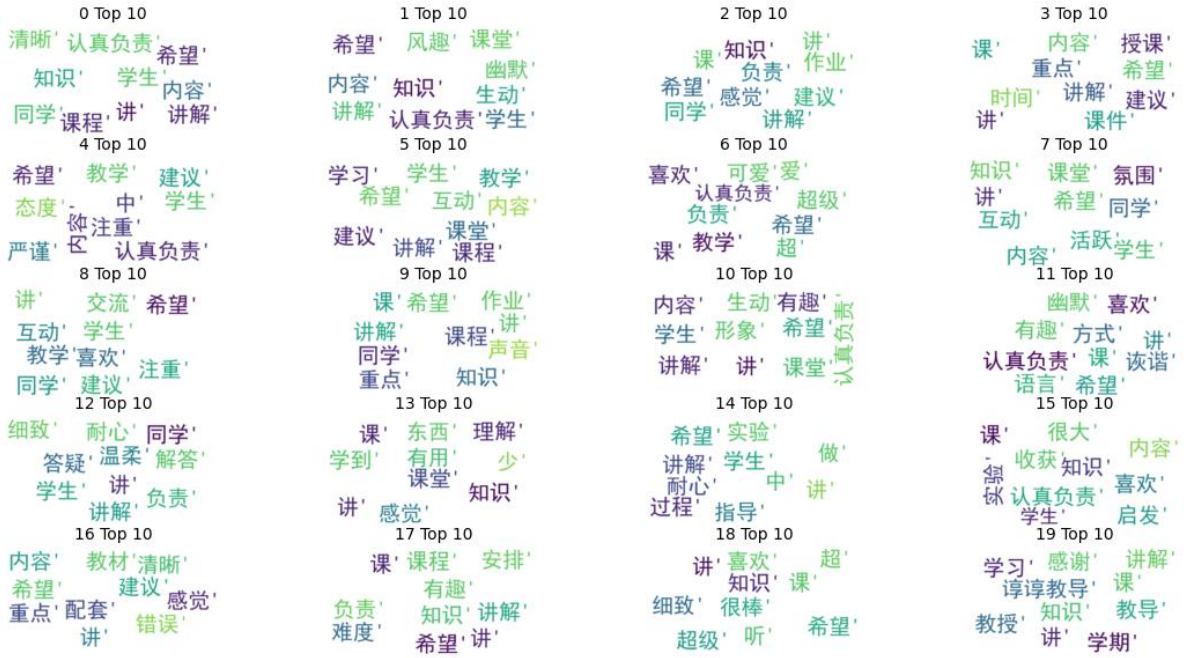


图 3.17 2017-2018 学年第二学期热点问题词云图

从图 3.17 可以看出在 2017-2018 学年第二学期，在第 8 类热点问题之中，有“交流”,“互动”,“注重”等的热点词汇，学生们希望在教课的过程之中可以多一些交流，注重课堂上的互动。在第 14 类，有“实验”,“指导”,“讲解”“耐心”等的热点词汇。说明学生们希望在实验职中多一些指导与讲解。提供了宝贵的建设性意见。在第 16 类热点问题，有“教材”,“配套”,“清晰”,“感觉”等热点词汇，可以看得出，在该学期，关于教材问题可能出现了一些错误 。

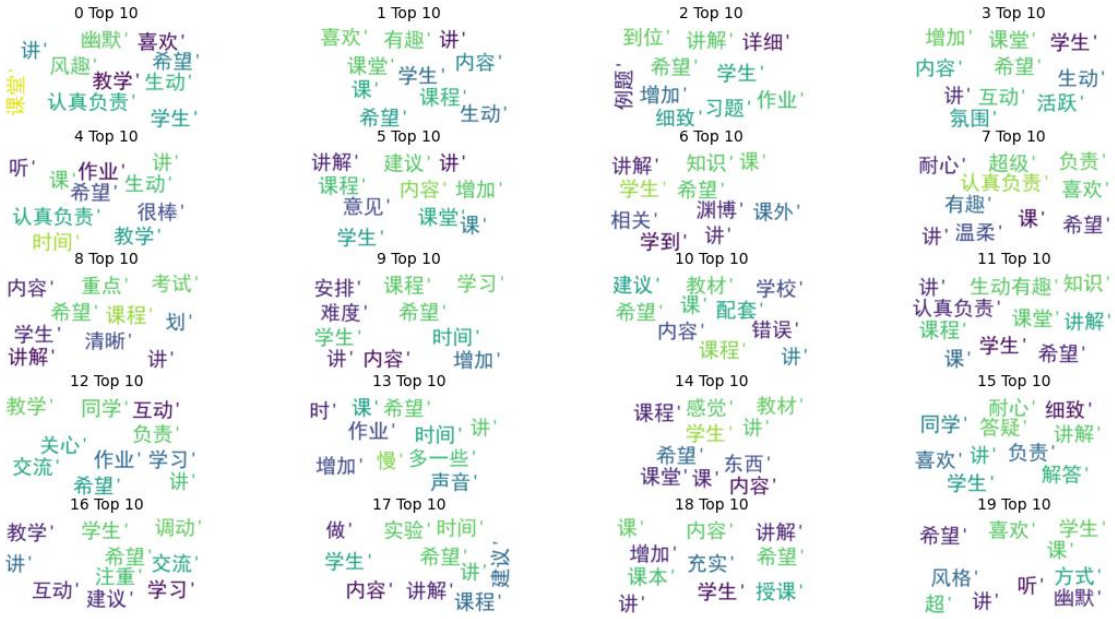


图 3.18 2018-2019 学年第一学期热点问题词云图

从图 3.18 可以看出在 2018-2019 学年第一学期，在第 8 类热点问题之中，有“重点”，“考试”，“课程”这些热点词汇。说明学生们还是希望在教学方面多进行课程重点知识的讲解，以及在讲课时应该要思路清晰。在第 10 类热点问题之中，有“教材”，“配套”，“内容”“错误”等的热点词汇，说明了，在教材方面上，学生们可能希望教材与所学教课内容一致。同时在 17 类热点问题之中，有“实验”，“指导”，“讲解”等的热点词汇。那么关于该学期，学生们希望关于一些实验课程，可以能够有一些老师们的讲解与指导，可以更好的来完成实验课程的内容。这项对以前是一个新的热点问题。同时也具有建设性的意见。

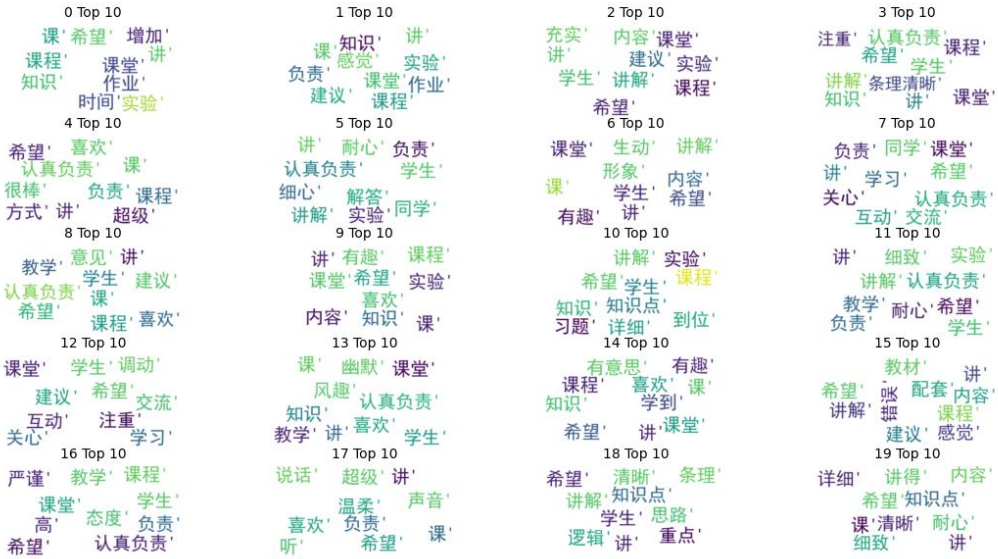


图 3.19 2018-2019 学年第二学期热点问题词云图

从图 3.19 可以看出在 2018-2019 学年第二学期，在第 15 类热点问题之中，有“教材”，“配套”，“内容”“错误”等的热点词汇，说明了，在教材方面上，学生们可能希望教材与所学教课内容一致。在第 18 类有“清晰”，“条理”，“讲解”等的热点词汇，那么说明学生们希望能够老师上课讲解有条理，思路清晰。同时希望重视重点知识的讲解。

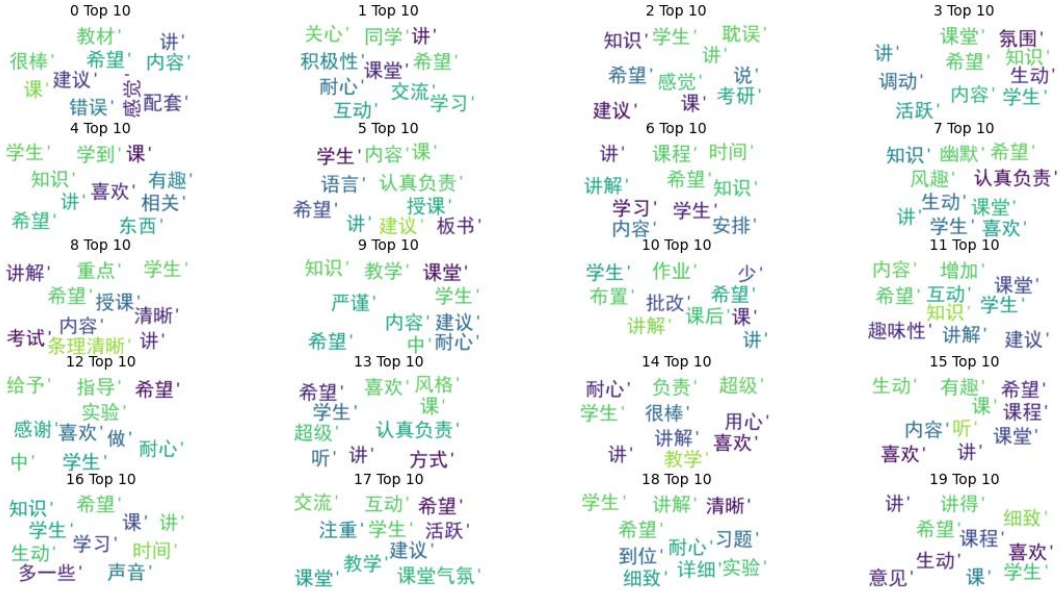


图 3.20 2019-2020 学年第一学期热点问题词云图

从图 3.20 可以看出在 2019-2020 学年第一学期，在第 0 类热点问题之中可以看出，该学期教材仍然有问题，在教材的和所教授的内容上面仍然有配套的错误。在第 8 类热点问题之中，有“重点”，“授课”，“清晰”等的热点词汇，可以看得出，同学们在上课的时候，还是多希望老师可以讲课条理清晰一些，注重重点知识的教授。

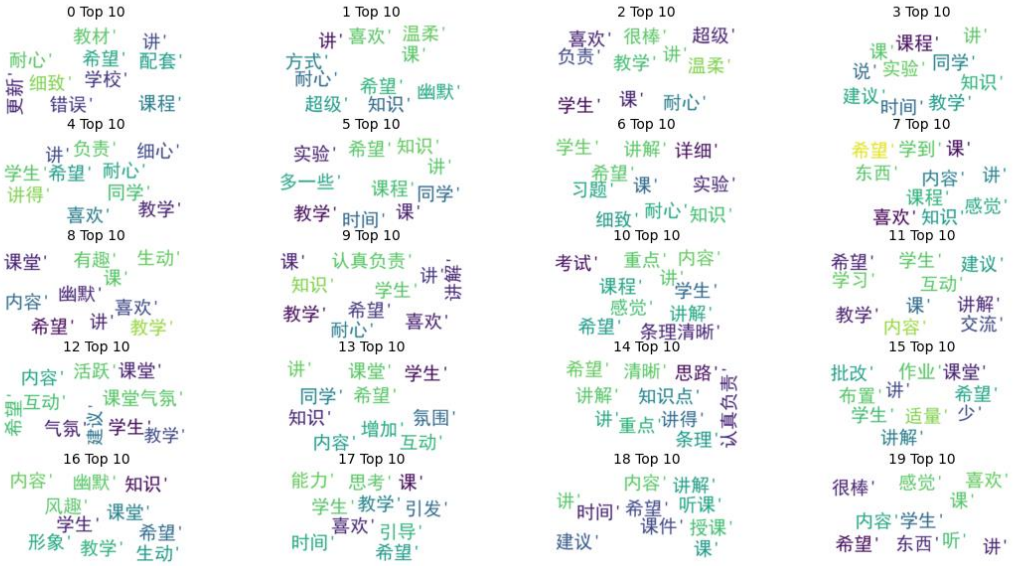


图 3.21 2019-2020 学年第二学期热点问题词云图

从图 3.21 可以看出在 2019-2020 学年第二学期，看的出在很多类热点问题之中，同学们的关注点还是主要在课堂上面，想要幽默，生动的课堂，同时也希望老师能够活跃课堂的气氛，注重重点知识的讲解，要思路清晰，同时要注重课堂上与同学们之间的交流。

3.4.5 整体数据结果分析

从这三年的数据来看，一直都有“幽默”，“耐心”，“负责”，“教材”，“重点”，“清晰”等的热点词汇。那么我们就可以看出，在教学之中，最重要还是要老师注重知识点的讲解，要让同学们觉得条理清晰，同时也要求老师们在教学之中要对同学们的问题进行耐心的解答。只有注重了教学之中的本质问题，多注重同学们反映出的真实的声音，才可以帮助教学质量的更进一步的提升。

3.4.6 算法运行时间效率分析

不同学期的执行时间 t(s)记录如表 3.1 所示。

表 3.1 所有学期算法的执行时间数据

学年学期	17-18-1	17-18-2	18-19-1	18-19-2	19-20-1	19-20-2
运行时间	571s	521s	703s	605s	638s	579s

不同学期的执行时间 t(s)如图 3.22 所示。

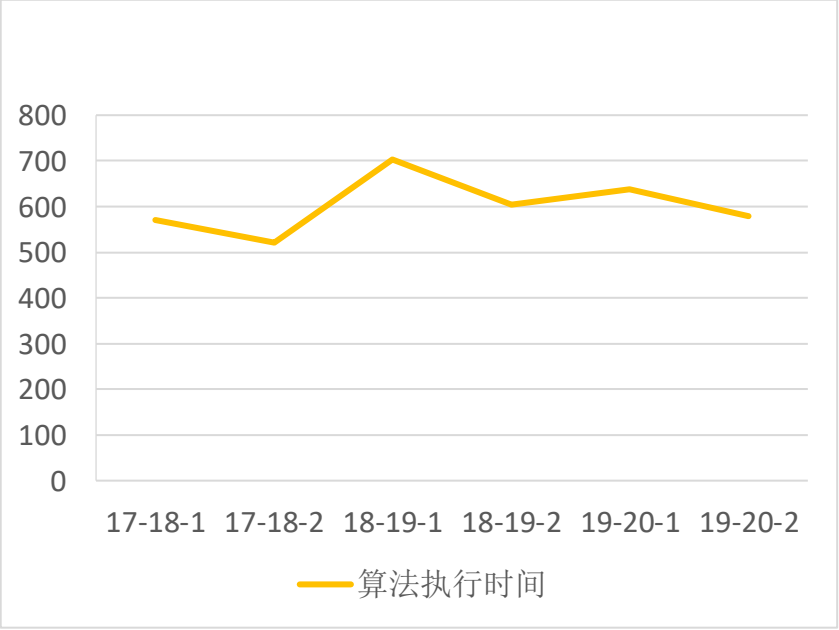


图 3.22 执行时间 t

从算法执行的时间可看，执行时间在 600s 左右浮动，对于一学期数十万量级的教学评估数据量来说，算是在合理的范围之内，在真实的服务器环境之中，运行速度会有大幅的提升，所以本课题之中的算法在时间复杂度和空间复杂度效率处在合理的范围之内。

第四章 系统设计与实现

4.1 系统需求分析

由于手动查找教学之中热点问题数据较为复杂麻烦，付出的时间成本巨大，因此本课题的热点问题发现系统运用的文本聚类算法进行教学评估无用数据的清洗，并且进行热点问题的发掘，并进一步找出关于热点问题的建设性意见，用于用户的进一步操作和分析。同时在性能方面也要做到安全，兼容，可靠以及可操作。

4.1.1 功能需求

教学评估热点问题发现系统为用户提供挖掘教学评估之中的热点问题的功能，用户可以按照自己的需求选择不同的学年学期的范围来查看热点问题，该系统需实现以下功能需求：

1. 教学评估热点问题结果发现

（1）选择相应学年学期并对教学评估数据进行热点问题发现，生成词云图。并且显示各类热点问题下的所有教学评估数据。

（2）对所查询的结果进行保存，方便下一次利用分析数据。

教学评估结果查询功能如图 4.1 所示。

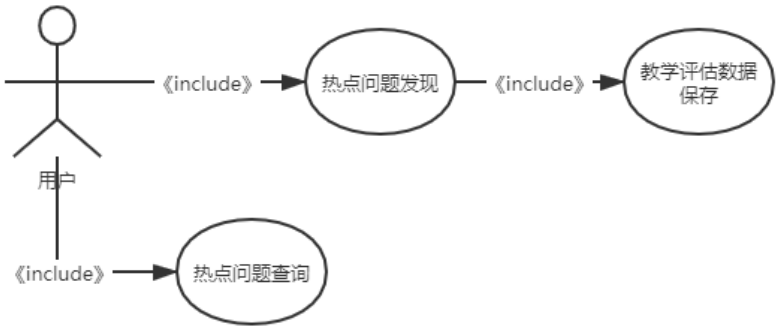


图 4.1 教学评估热点问题结果发现用例图

4.1.2 性能需求

为了确保教学评估热点发现系统能够良好运行和使用，本系统的性能需求如下：

（1）安全性需求

本课题中的教学评估热点问题发现系统应该具有一定容错能力，保证系统运行的可靠以及文本聚类过程的准确。保证数据库数据的完整以及正确，为用户提供一个安全的热点问题发现系统，允许用户进一步的分析数据。

（2）运行速度需求

教学评估热点问题发现系统的运行时间应该一个合理的范围之内，系统运行时间不应过长，所以需要考虑时间复杂度较小的热点问题发现算法。

（3）实用性需求

热点问题发现系统的开发应该符合学校教务处的实际需求，符合学校教务处的实际情况，可以真正意义上的帮助学校发现教学之中的热点问题，发现具有建设性的意见，方便学校教务处对教学评估数据进行更近一步的分析与使用，从而为推动学校教学方面的改革。

（4）可扩展性需求

本课题之中教学评估热点问题发现系统的开发需要符合学校教务处的框架，代码实现的流程与学校教务处的流程相一致，同时需要预留 python 的 WebAPI 接口方便用户调用后端的核心算法，同时代码的格式要清晰，规范，核心的地方需要标明注释，为接下来的进一步开发提供一个良好的基础。

（4）易用性需求

关于用户界面的设计应该简洁，方便用户的使用，同时在结果展示方面，也多应用图片的形式进行展示，使得结果更加直观。

4.2 系统设计

4.2.1 系统框架设计

根据需求分析，本课题的热点问题发现系统围绕热点发现任务为核心，教务人员可以对发现后的热点问题进行相关的信息查看。

本科教学评估热点问题发现系统拟采用三层 B/S 架构，网页后端代码用 C#实现。并调用 WebAPI 服务，并传送返回值，实现了表示层与服务层的连接。核心算法代码使用 python 语言实现。在确定好参数和返回值之后封装成为 WebAPI 接口，通过相关的 SQL 语句访问数据库获得相应的数据，这样就连接起了数据层和服务层。

系统分为数据层、服务层和表示层一共三层。数据层存储系统用链接到本地的 SQLserver 数据库。服务层作为中间层，包括 Web API 服务。Web API 服务通过 Httprequest 接收 B/S 客户端的请求，根据业务逻辑进行相应的处理，最后通过 Httprequest 响应客户端的请求并返回最终结果。客户端访问数据层的本地数据库来获取数据。B/S 客户端在表示层通过用户的操作来响应相应的事件的发生。这样就做到了分层的功能，达到了封装效果的同时可以使得各层之间相互独立。如图 4.2 所示。

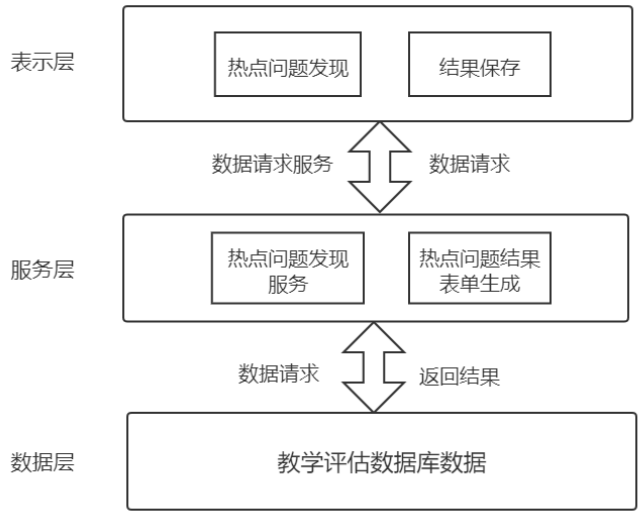


图 4.2 系统总体架构

4.2.2 系统功能设计

根据系统的需求分析，系统的主要功能模块可分为数据热点问题发现模块。

(1) 数据清洗模块

数据清洗模块对数据进行去重，去空以及删除无用的数据，在后端执行，不会在前端展示。数据清洗模块如图 4.3 所示。

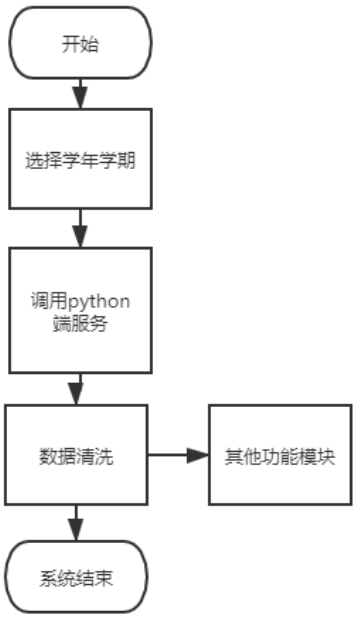


图 4.3 数据清洗模块

(2) 数据热点问题发现模块

数据热点问题发现模块实现发现教学评估之中热点问题，用户可以选择想选择的学年和学期，点击按钮响应事件，就可以执行对应的 SQL 语句在数据库之中进行相应的查询，

并将参数通过接口传给 python 写好后端的算法，对数据进行计算处理后，得到所选学年学期的每一类热点问题的所有教学评估数据，显示在表格之中。数据热点问题发现模块如图 4.4 所示。

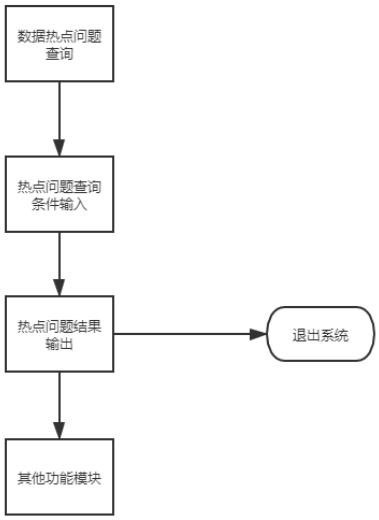


图 4.4 数据热点问题发现模块

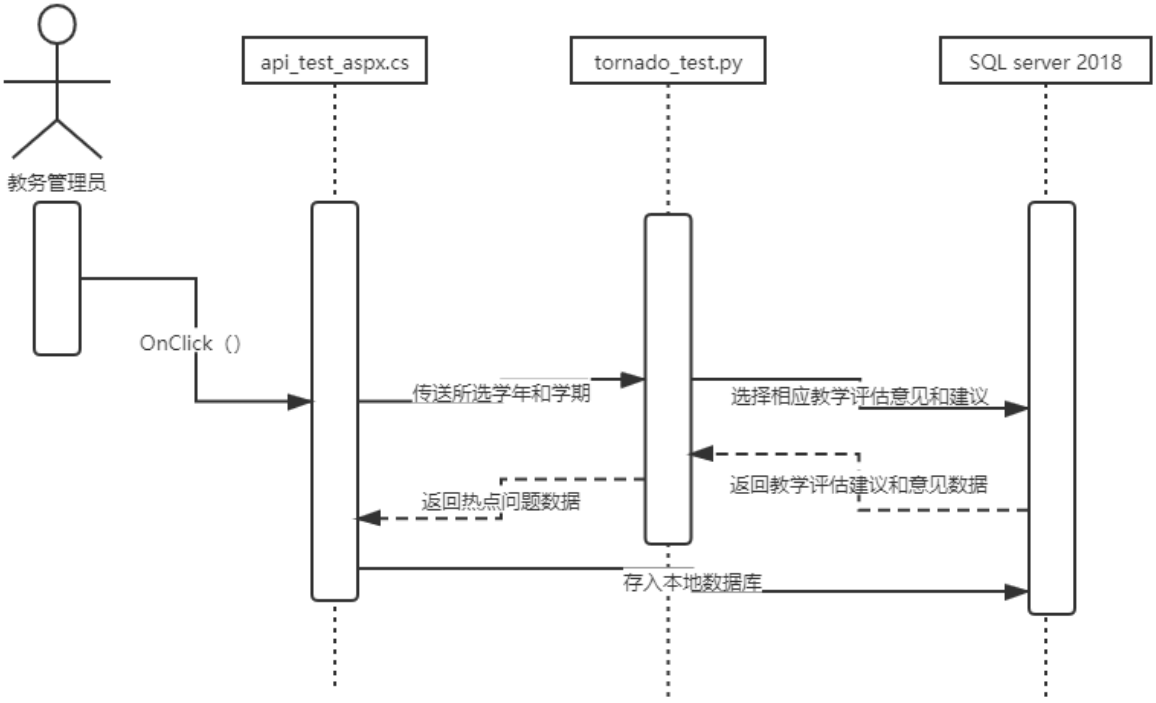


图 4.5 数据热点问题发现流程图

(3) 热点问题查看模块

在发现了所选学年学期的热点问题之后，可以对之前所发现的结果进行查看。流程如图 4.6 所示。

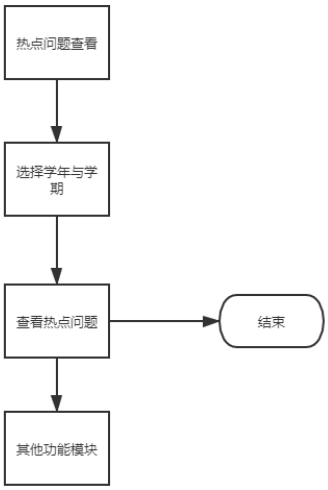


图 4.6 热点问题数据查看模块

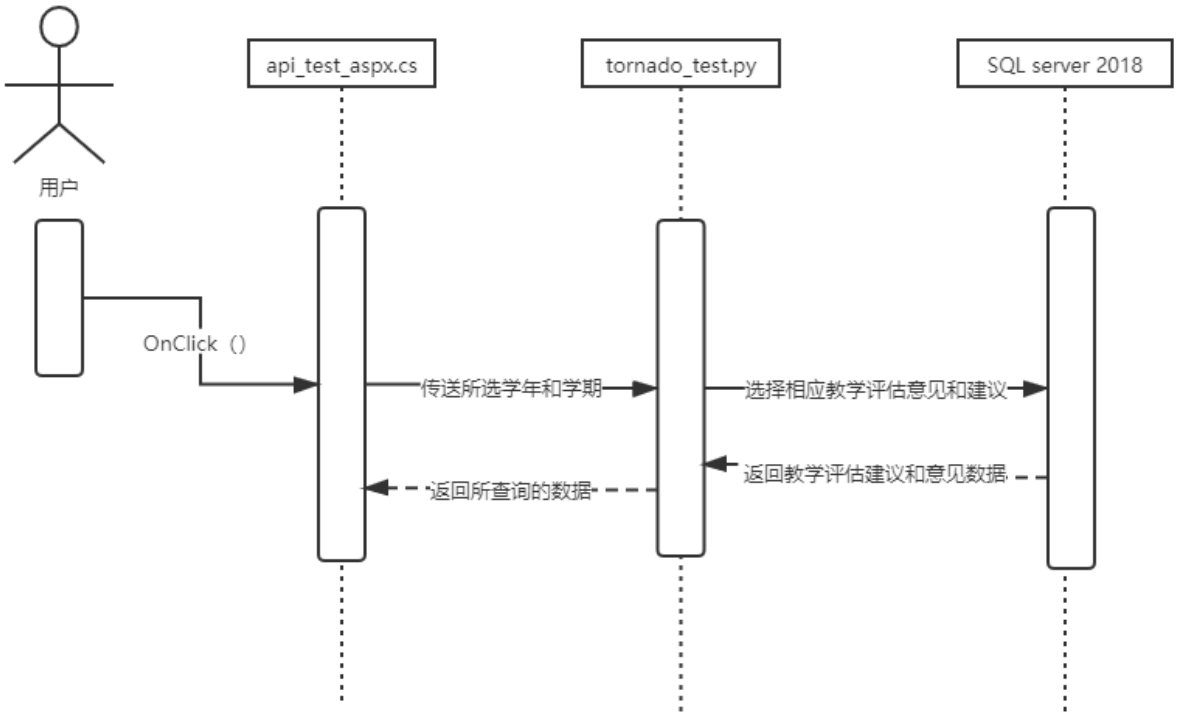


图 4.7 热点问题数据查看流程图

4.2.3 系统类图

PG_JGB 中记录所有学生的评教结果，是文本聚类的主体。PG_RDWT 中记录所发现的热点问题。PG_RDWTFX 中记录教务员所发现热点问题。三个类图的关系如图 4.8 所示。

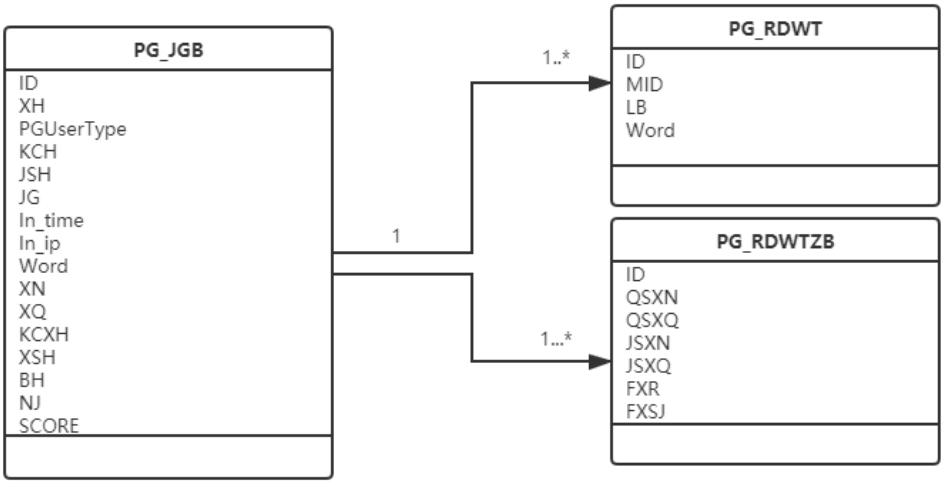


图 4.8 系统类图

4.2.4 数据库设计

本科教学评估数据热点发现系统数据库关键表的详细设计如下所示。PG_JGB 中记录所有学生的评教结果，是文本聚类的主体。 如表 4.1 所示。

表 4.1 评估结果表 PG_JGB

列名	数据类型	允许 Null 值	中文意义
ID	int	N	序号（自动增长）
XH	nvarchar(20)	N	学号/专家号
PGUserType	nvarchar(10)	N	评估人员类别
KCH	nvarchar(20)	N	课程号
JSH	nvarchar(20)	N	教师号
JG	nvarchar(4000)	N	结果
In_time	Datetime	N	评估时间
In_ip	nvarchar(50)	N	评估机器 IP
Word	nvarchar(4000)	N	评语
XN	varchar(50)	N	学年
XQ	varchar(50)	N	学期
KCXH	varchar(50)	N	课程序号
XSH	varchar(10)	Y	学生号
BH	varchar(20)	Y	班号
NJ	varchar(10)	N	年级
SCORE	decimal(18,2)	Y	总评成绩

PG_RDWTZB 中记录教务员所发现热点问题，有所选择的时间范围，分析人和分析时间等的属性来记录，详细设计如表 4.2 所示。

表 4.2 热点问题发现表 PG_RDWTZB

列名	数据类型	允许 Null 值	中文意义
ID	int	N	主表编号
QSXN	varchar(10)	N	起始学年
QSXQ	varchar(2)	N	起始学期
JSXN	varchar(10)	N	结束学年
JSXQ	varchar(2)	N	结束学期
FXR	varchar(50)	N	分析人
FXSJ	Datetime	N	分析时间

PG_RDWT 中记录所发现的热点问题。 其中类别为该类热点问题下最高频的词汇。例如“课堂”为最高频词汇，则为“课堂”类。详细设计为表 4.3 所示。

表 4.3 热点问题表 PG_RDWT

列名	数据类型	允许 Null 值	中文意义
ID	int	N	序号（自动增长）
MID	int	N	主表编号
LB	varchar(50)	N	类别
Word	nvarchar(4000)	N	热点问题

4.3 系统实现

4.3.1 系统开发环境

（1）Visual Studio 2017

Visual Studio 2017 是一款非常全能的开发软件，它支持多种语言和项目的开发环境，在开发的过程之中，Visual Studio 2017 为用户提供很多可以自行安装的插件，方便一些所必须完成的项目的管理。同时 Visual Studio 支持不同编程语言之间的串联，相关的 Web 的接口的调用，做到了整合一体化。

（2）PyCharm

PyCharm 是一种 Python 的编译器，是一款非常实用的 python 语言开发工具，它为用户提供了很多实用的功能，比如关键词的高亮，代码的跳转，多种框架的项目管理，比如 flask 框架以及 Django 框架。

（3）SQL Server 2018

微软的 SQL Server 2018 是一个全面的数据库平台，同时，微软还有 SQL Server 2018 management 可以进行数据库数据的可视化操作。拥有可靠的数据存储功能，使用户完全可以建立高性能的数据程序。可以满足学校教务处对于数据的要求。

4.3.2 基础功能实现

（1） 通过 python flask 封装 Web Service 接口

通过 python 所带 flask 框架包，将整体运行的代码封装成为在服务器端运行，其中核心的框架代码 ：

```
#创建 flask 对象
app = Flask(__name__)
#创建路由'/'
@app.route('/',methods=['POST'])
def home():
    if request.method == 'POST':
            receiveData = request.data.decode('utf-8')  # 为了兼容中文
            para = str(receiveData)
            print(para)
    return sum_text

      #实现整体热点问题发现的代码
#当用户请求'/'资源时，回传 sum_text
#启动 flask，并设定端口为 5000
app.run(port = 5000)
```

（2） 通过 C# HttpWebRequest 调用 Web Service

响应 click 事件后，之后通过 ip 地址创建 url，同时设置 HttpWebReques 的相关参数，其中比较重要的是 wRequest.Timeout= Timeout.Infinite;保证了请求时永远不超时，可以允许我们传输较大的数据文件。读取前端所选择的学年学期，保存成为 json 格式的字符串，并以字节的方式写入流，之后请求回复，从 python 的代码之中的到最终的结果。并最终打印到网页前端。

主要流程如图 4.9 所示。

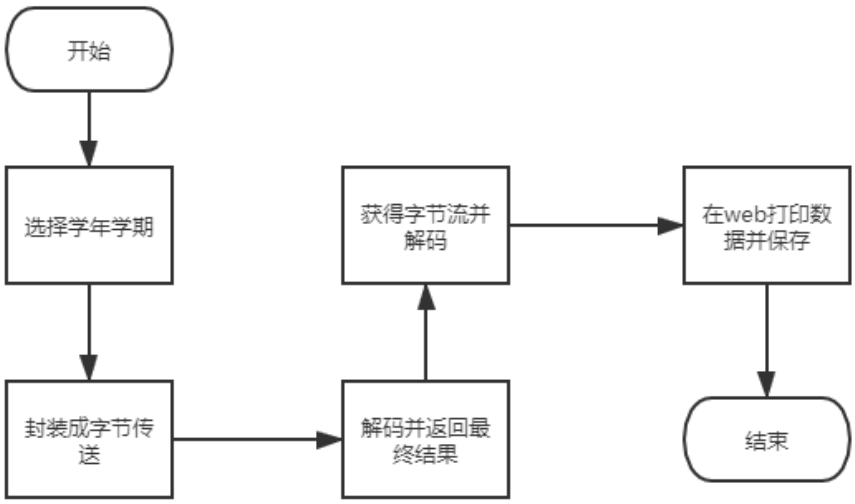


图 4.9 数据热点问题发现模块

核心代码如下：

```
protected void Button1_Click(object sender, EventArgs e)
        {
            string XN = DropDownList1.SelectedValue;
            string XQ = DropDownList2.SelectedValue;
            string jsonParams = "#" + XN + "#" + XQ + "#";
            string url = "http://127.0.0.1:5000";
            HttpWebRequest wRequest = (HttpWebRequest)WebRequest.Create(url);
            wRequest.Timeout = Timeout.Infinite;//放到一开始声明变量的时候成
功了
            wRequest.ServicePoint.Expect100Continue = false;
            wRequest.KeepAlive = true;
            wRequest.Method = "POST";
            wRequest.ContentType = "application/json";
            string paraUrlCoded =
jsonParams;//System.Web.HttpUtility.UrlEncode(jsonParas);
            byte[] payload;
            //将Json字符串转化为字节
            payload = System.Text.Encoding.UTF8.GetBytes(paraUrlCoded);
            //设置请求的ContentLength
            wRequest.ContentLength = payload.Length;
            //发送请求，获得请求流
            Stream writer;
            try
            {
                writer = wRequest.GetRequestStream();//获取用于写入请求数据的
Stream对象
            }
            catch (Exception)
            {
```

```
                writer = null;
                int i = -1;
            }
            //将请求参数写入流
            writer.Write(payload, 0, payload.Length);
            writer.Close();//关闭请求
            int k = 0;
            HttpWebResponse wResponse =
(HttpWebResponse)wRequest.GetResponse();
            Stream stream = wResponse.GetResponseStream();
            StreamReader reader = new StreamReader(stream,
System.Text.Encoding.UTF8);
            string str = reader.ReadToEnd();    //url返回的值
            reader.Close();
            wResponse.Close();
            }
```

4.3.3 系统运行结果

(1)系统主界面

如图 4.10 所示。

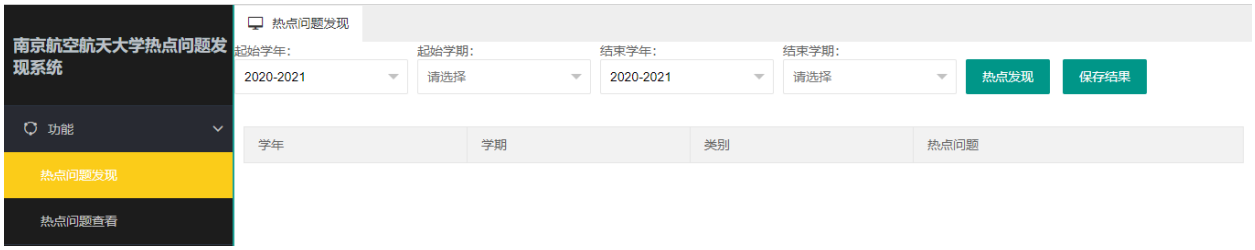


图 4.10 系统主界面

(2)教学评估热点问题结果查询

用户选择学年学期后，点击查询按钮，就可以按照类别得到该学期下的所有热点问题的教学评估数据。如图 4.11 和 4.12 所示。

图 4.11 2019-2020 学年第一学期热点问题查询结果 1 讲解类

图 4.12 2019-2020 学年第一学期热点问题查询结果 2 希望类

第五章 总结与展望

5.1 工作总结

从本科教学评估数据之中提取教学之中的热点问题并对其进行分析，可以真实的反映出学生们所关注的教学之中的热点问题，有助于推动教学方面的改革，更直观的看见学生们的意见。本文对本科教学评估数据进行数据预处理，文本聚类，并最终将热点问题展现出来，同时进行了相应的系统的开发。总结工作如下：

（1）阐明了本课题之中聚类算法的相关背景，意义和研究现状。同时也说明了本篇论文整体的组织架构。

（2）学习并研究了课题所需要的算法原理，介绍了数据预处理，空间向量模型，TF-IDF 权重，文本相似度测量，K-means 算法的原理等的与本课题相关的算法原理。

（3）实现了文本聚类，文本挖掘的功能，实现了热点问题发现系统的开发与维护。

（4）学习 Web 平台 B/S 三层架构的开发，生成 WebAPI 服务的方法，等系统开发所需知识，根据学校教务处的要求，完成了整个系统功能的开发。

（5）实现了结果直观的展示，同时也经过了系统的测试，结果和速度均满足运行的要求。

5.2 工作展望

本文实现了对本科教学评估数据进行热点问题发现的实现，基于 flask 的 API 框架去进行接口的开发，并且为了使其能够更好地融入教学评估系统之中，使用了 C#的 WebApplication 框架去调用 python 的 API 接口。在学习的过程之中遇到了不少的问题与困难，好在自己一步一步的努力进行克服，做到了用户真正可以发现每一学年学期之中的教学评估之中的热点问题。同时也提升了自己系统的开发能力，学习到了调用 API 接口的能力以及 python 语言的开发能力。但是，目前关于算法的时间复杂度还是很大，需要相处更好地方法进行时间上的更优化。同时如何关于每类热点问题之中的评论，如何希望本文所实现的结果可以帮助到发现教学之中存在的热点问题，从而进一步帮助到教学改革，提升本科教学质量。

参考文献

[1]吴启明,易云飞.文本聚类综述[J].河池学院学报,2008(02):86-91.

[2]李海磊,杨文忠,李东昊,温杰彬,钱芸芸.基于特征融合的K-means微博话题发现模型[J].电子技术应用,2020,46(04):24-28+33.

[3]冯俐.中文分词技术综述[J].现代计算机(专业版),2018(34):17-20.

[4]彭敏,黄佳佳,朱佳晖,黄济民,刘纪平.基于频繁项集的海量短文本聚类与主题抽取[J].计算机研究与发展,2015,52(09):1941-1953.

[5]王俊丰,贾晓霞,李志强.基于K-means算法改进的短文本聚类研究与实现[J].信息技术,2019,43(12):76-80.

[6]李伟. 基于频繁词集词共现网络的短文本聚类方法[D].北京交通大学,2016.

[7]李岩. 基于深度学习的短文本分析与计算方法研究[D].北京科技大学,2016.

[8]雷鸣. 基于文本挖掘技术的社会热点分析[D].浙江工商大学,2019.

[9]李廷进. 基于主题模型的文本聚类研究与应用[D].山西大学,2020.

[10]马英杰.浅谈数据挖掘技术在教学实践中的应用[J].课程教育研究,2018(38):238.

[11]谢代邑. 基于数据挖掘的福建师范大学教学质量评估系统的设计与实现[D].电子科技大学,2013.

[12]Khaled Abdalgader (November 5th 2018). Centroid-Based Lexical Clustering, Recent Applications in Data Clustering, Harun Pirim, IntechOpen, DOI: 10.5772/intechopen.75433. Available from: https://www.intechopen.com/books/recent-applications-in-data-clustering/centroid-based-lexical-clustering

[13]胡章平. 基于模糊综合评判的教师教学质量评估系统的设计与实现[D].重庆大学,2006.

[14]Tu Shifen,Yang Bo. Research on sentiment classification of micro-blog short text based on topic clustering[J]. Journal of Physics: Conference Series,2021,1827(1).

[15]Di Wu,Ruixin Yang,Chao Shen. Sentiment word co-occurrence and knowledge pair feature extraction based LDA short text clustering algorithm[J]. Journal of Intelligent Information Systems,2020(prepublish).

[16]刘洁晶. 数据挖掘技术在教学评估预测系统中的应用研究[D].河北大学,2012.

[17]Kongwudhikunakorn Supavit,Waiyamai Kitsana. Combining Distributed Word Representation and Document Distance for Short Text Document Clustering[J]. Journal of Information Processing Systems,2020,16(2).

[18]李旭军.数据挖掘技术在高校教学质量评估中的应用[J].重庆科技学院学报(社会科学版),2012(01):177-179.

[19]Junyang Chen,Zhiguo Gong,Weiwen Liu. A Dirichlet process biterm-based mixture model for short text stream clustering[J]. Applied Intelligence,2020(prepublish).

[20]Wang, Ai,Gao, Xuedong. Multifunctional Product Marketing Using Social Media Based on the Variable-Scale Clustering[J]. Tehnički vjesnik,2019.

[21]姜允志,宋新红.基于短文本聚类的学生评教方案设计[J].教育教学论坛,2020(44):346-347.

致　　谢

这次毕业设计能够完成，毕业论文可以完成，需要感谢很多我身边的老师，朋友，同学们，当然，也要在这里感谢自己，感谢自己努力坚持学习，不会轻易就选择放弃。遇到了困难也学着尝试努力克服。

首先，在这里我要先对我的毕业设计的指导老师，谢强老师表达我由衷的感谢，感谢老师在我遇到困难的时候给与我指导与帮助。一起共同完成了整个毕业设计的项目。谢强老师也是我编译原理课程以及.NET web 开发课程的教课老师。从课程之中也收到了不少专业知识，帮助了我毕业设计的学习与开发。

其次，我也要感谢我的家人，谢谢他们一直以来都在尊重和支持我的决定。所以我也有机会一直以来都可以做自己想做的事情。我知道这样很不容易，或许有的时候结果也会令他们感到失望，他们也许也有过担心和迟疑，但却无一例外的仍然选择继续尊重并且支持我的决定。能得到家人们的尊重和支持，这在这个世界上是很难得的经历与体验。所以我要好好珍惜并且不算努力。完成自己想完成的事情，不辜负他们的期待。同时也要感谢他们让我懂得要选择保持坚强和独立，这是很可贵的。

回想起大学四年的生活，关于学习的经历，确实也有过无数个和 bug 有来有回的夜晚，有过不少的痛苦的经历，每一次的学期末都是在和期末考试与课程设计之间努力的挣扎，所以要感谢自己没有妥协，尽自己最大能力的去面对每个任务，不管最后的结果如何，中间经历的过程，都将是我人生宝贵的经历。抛去学业上的繁忙，我确实在大学期间也结实了很多真心的朋友。相信以后回忆起来，不会是和 bug 斗争的夜晚，而是回想起和朋友们一起享受过的快乐、短暂且宝贵的时光。与他们相处了四年短暂且又美好的大学生活，是我本人这辈子的宝贵经历，四年之中，有过惊喜，快乐，感动，当然也曾有过迷茫，痛苦和焦虑。不过希望未来回首这段经历，留下的皆是美好的回忆。愿你们永远学会坚强，努力保持纯粹。未来大家也都会有自己不同的选择。希望你们前程似锦，同时自己也要好好努力，让自己得到自己想要的，成为自己想成为的人。

最后，祝母校越来越优秀，为祖国建设培养出越来越多的计算机人才。推动社会主义现代化强国的建设与发展，让祖国成为世界之中真正意义上强大的国家，只有真正意义上的实力变强了，在可以让祖国在这个世界上站稳脚跟。为人类科技的进步做出我们中国人的努力和贡献。