学号 2015302580330

密级 公 开

武汉大学本科毕业论文

食品安全网络公开数据采集技术研究

院（系）名 称：计算机学院

专 业 名 称 ：软件工程

学 生 姓 名 ：文晴曼

指 导 教 师 ：朱卫平 副教授

二○一九年五月

**BACHELOR'S DEGREE THESIS**

**OF WUHAN UNIVERSITY**

**Research on Food Safety Open Data Collection Technology**

College ： School of Computer Science

Major ：Software Engineering

Name ：Qingman Wen

Supervisor ：Dr. Weiping Zhu

May 2019

**郑 重 声 明**

本人呈交的学位论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本学位论文的知识产权归属于培养单位。

本人签名： 日期：

摘 要

随着生活水平的提高，人们越来越关注食品安全问题。近几年发生的一些食品安全恶劣事件对社会造成了严重的不良影响，人们迫切地需要改善当前食品安全的情况。本文采用计算机技术来对食品安全领域进行研究。

由于网络上关于食品安全的数据格式各异，我们很难直接对其进行分析和处理。本文通过采集网络平台上各类型有关食品安全的数据，对格式复杂的数据例如音视频、图片等进行转化，从中提取出有效信息，使得人们能够有效利用各类型数据进行分析。主要研究内容可以分为三部分，对网络公开数据的爬取，对各类型数据的处理，对文本内容的分析。首先，从网络平台上获取食品安全话题下的不同类型数据，这些网络平台包括了文本、音视频网站等。然后，对爬取到的数据进行预处理，从音视频数据中提取文字内容，从图片中识别实物，从而使计算机能够快速识别这些复杂类型数据所传递的信息。最后，对文本数据进行了研究，利用文本分类技术来进行话题判别，剔除不属于食品安全话题的数据；利用文本聚类技术对新闻报道进行聚类，直观表现一段时间内关于食品安全的具体话题类别。

关键词：食品安全；爬虫；多媒体数据处理；文本处理

**ABSTRACT**

With the improvement of living standards, people are paying more and more attention to food safety issues. Recent years, some adverse food safety incidents have had serious adverse effects on society, and there is an urgent need to improve the current food safety situation. This paper uses computer technology to conduct research on the field of food safety.

This paper studies the collection of food safety public data on the network platform. The main research content can be divided into three parts, the crawling of the public data of the network, the processing of various types of data, and the analysis of the text. First, different types of data under the food safety topic are obtained from the network. These network platforms include text, audio and video websites. Then, the crawled data is preprocessed, the text content is extracted from the audio and video data, and the real object is identified from the image, so that the computer can quickly identify the information transmitted by the complex type data. Finally, the text data is studied. We used text classification technology to do topic discrimination, eliminating data that is not about food safety. Then we used text clustering technology to cluster news reports to show specific topics about food safety in a period of time.

**Key words:** food safety; web crawler; multimedia data processing; text processing

目 录

第一章 绪论 1

1.1 研究背景与意义 1

1.2 国内外研究现状 2

1.3 本文研究内容与论文结构 4

1.4 本章小结 5

第二章 相关技术 6

2.1 网络爬虫技术 6

2.1.1 爬虫分类 6

2.1.2 爬虫工具 7

2.2 非结构化数据处理 8

2.2.1 视频文字提取 9

2.2.2 音频文字提取 9

2.3 文本处理技术 10

2.3.1 文本预处理 10

2.3.2 文本表示 10

2.3.3 文本分类 12

2.3.4 文本聚类 12

2.4 本章小结 13

第三章 数据采集 14

3.1 基于Scrapy的爬虫 14

3.2 MongoDB存储数据 14

3.3 多媒体数据处理 15

3.3.1 视频文字提取 15

3.3.2 音频识别 17

3.3.3 图片实体识别 18

3.4 实验与分析 19

3.5 本章小结 23

第四章 利用文本分类进行数据过滤 24

4.1 文本预处理 24

4.2 基于文本分类的主题判别 24

4.2.1 TextRank提取关键词 25

4.2.2 文本向量化 25

4.2.3 构建分类器 27

4.3 实验与分析 28

4.4 本章小结 31

第五章 话题提取 32

5.1 K-means文本聚类分析 32

5.1.1 K-means算法 32

5.1.2 利用K-means进行文本聚类 32

5.2 LDA分析话题 34

5.3 实验与分析 36

5.4 本章小结 38

第六章 结语 39

6.1 总结 39

6.2 展望 39

参考文献 41

致谢 43

# 第一章 绪论

本章主要介绍食品安全网络公开数据采集技术的研究背景与意义、国内外研究现状和本论文的总体结构。

## 1.1 研究背景与意义

随着人民生活水平的提高，“食品安全”话题在近几年越来越受到人们的关注。“毒大米”、“假酒”、“毒奶粉”等恶劣事件还未淡出人们的视野，近段时间又有多家知名餐厅甚至学校食堂被曝后厨脏乱差。这些问题已经严重危害了人们的生活。因此政府部门需要严格监管食品安全，在发现问题食品后及时控制局面，将其对人民和社会造成的不良影响降到最低。普通民众也需要及时掌握关于有害食品的信息，避免误食有害食品，危害健康。而在大数据时代，人们在网络上的交互十分频繁，信息的传播和扩散变得更加迅速。每天有数以万计的数据曝光在我们面前，我们无法通过传统数据处理软件来对这些数据进行管理、获取以及处理。这些数据通常被分为两类，非结构化数据和结构化数据。结构化数据被存储于数据库中，有明确的数据格式。而非结构化数据没有完整的数据格式，包含了文本、图片、html网页、音视频等数据。Amir Gandomi等提到非结构化数据占据了大数据的95%[1]。

从这些格式各异的网络数据中获取食品安全领域的信息是我们需要解决的问题。网络平台上大多都是不便于存储和处理的非结构化数据，包含文本、图片、音视频数据、html网页等。由于非结构化数据包含的数据格式有很多，因此目前研究者们在研究非结构化数据时将范围缩小，通常只针对其中的一个方面进行研究。Amir Gandomi等人认为大数据分析包含了文本分析、音频分析、视频分析和社交媒体分析这几个方面。他们提到文本分析包含信息提取、文本摘要自动生成、自动问答系统和情感分析；音频分析是从非结构化音频数据中分析和提取信息；视频分析是从视频流中提取出有价值的信息；社交媒体分析是对社交媒体中的内容或实体进行分析[1]。Gharehchopogh F S A 等人则主要对文本分析进行了研究，他们使用文本挖掘技术将文本结构化[2]。杨桥则是对如何存储非结构化数据进行了研究[3]。

网络平台上的数据量庞大，人们很难在较短时间内获取到与某个话题相关的大量数据。而且，由于网络中的数据有多种格式，即使获取到了也很难直接进行分析，例如音视频通常是以流的方式存在，我们只有播放音视频才能了解其传递的语义信息，而播放的过程会消耗巨大的精力和时间，当我们有许多音视频待分析的时候，这种方式就不可行。另外，当我们获取到大量数据的时候，其中可能会混入一些不属于食品安全话题下的数据，手动的从成百上千的数据中剔除那些垃圾数据将会造成人力资源的浪费并消耗大量时间。因此本文对网络平台食品安全话题下数据的采集技术进行了研究，采集数据并进行统一处理。

本文的研究对人们快速从网络平台上获取食品安全话题的数据有着重要意义，使得利用机器代替人工进行音视频、图片的分析成为了可能，并且对政府部门快速了解一段时间内的食品安全热点话题并及时控制局面能够发挥很大的作用。

## 1.2 国内外研究现状

食品安全话题不论在什么年代都是人们关注的重点，互联网的发展使得人们可以利用计算机技术来对这一领域进行研究。各国计算机领域的研究者们往往从几个不同方面对食品安全领域进行了研究，其中包括了对食品安全事件的监测，对食品安全供应链的追溯以及监管等。

在国内，在对食品安全事件的监测方面，文献[4-8]用不同方法研究了对食品安全事件新闻的话题监测。王翠首先从网络上采集大量关于食品安全的新闻报道，然后在Hadoop云计算平台上，利用文本聚类算法对这些新闻报道聚类，聚类结果能够反映出不同食品安全话题的分布[4]。吴强强使用了主题爬虫技术来获取数据，实现了能根据特定主题从网络上爬取数据的系统，他比较了多种聚类算法，最后选用基于向量空间模型的改进的single-pass聚类算法来对数据进行处理，从而有效地从大量数据中提取出关键话题[5]。戴龙龙为提高话题的时效性，利用Storm分布式框架对single-pass算法进行改进，并提出一种建立食品决策树的方法，先确定每个文本所属的食品类别，再聚类文本，大大提高了聚类的速度。最后他将新闻报道、检测数据、品控数据和法律法规数据进行了匹配，使得系统的监控更具有参考价值[6]。杨柳讨论了不同长度文本的特性，对话题的关键词识别和特征提取方面进行了研究，并提出一种图结构模型来实现话题聚类和突发话题检测[7]。刘金硕等提出使用LDA模型对文本进行建模得到每篇文档的向量，然后利用K-means算法聚类，从而从食品安全数据中提取话题[8]。文献[9,10]研究了食品安全的追溯。其中汪历伟建立了食品供应链的数学模型，将食品生产、加工企业的数据集成到系统中，实现了一个基于SOA服务模式的食品安全追溯平台，让消费者和监管部门能够对食品的生产加工以及配送销售各个步骤进行追溯[9]。而林延昌认为近年来兴起的区块链技术可以有效解决食品追溯系统存在的普遍问题，他尝试将区块链技术运用到食品追溯领域，设计了一个基于区块链的牛肉安全追溯模型和策略[10]。通过上述分析我们可以看到，目前人们多使用文本数据对食品安全进行研究，而对其他类型数据的研究较少。

国外计算机领域的学者们针对食品安全领域的研究较少，Kate K讨论了新加坡国家环境局构建的FoodSIS系统的实现和构造，该系统是食品安全端到端网络信息采集系统，通过机器学习技术来识别和排序相关内容[12]。[13-14]介绍了大数据在食品安全领域的应用。Marvin H J P等人从食品安全数据收集、数据存储和传输、数据分析几个方面描述了大数据在食品安全中发挥的作用。他认为食品安全数据来源主要有数据库、互联网、基因库、社交媒体等，利用这些平台的数据可以对食品安全进行研究[13]。Ahearn M C等人讨论了大数据在食品安全供应链指导决策中发挥的作用[14]。Meyer C等人认为从私人组织和公共机构得到的数据往往具有滞后性，他们选取Twitter和Wikipedia作为数据源，首先利用基于规则和关键词的过滤器来过滤噪声数据，然后分析了Wikipedia的浏览量并发现浏览量数据可以作为食品预警信号，最后针对Twitter中的数据进行研究，发现当某一食物疾病传播时人们通常会发表一些与该疾病有关的症状描述言论。他们的研究让食品安全危机监测的时效性得到了提高[15]。Singh A等人以牛肉供应链为例进行研究，他们提到人们在Twitter中发表的言论往往能反映牛肉的质量问题。他们用文本挖掘和自然语言处理方法来分析Twitter数据的内容，利用SVM算法对数据进行情感分析，利用层次聚类算法分析不同情感分类下数据的关键词，从而可以得到食品供应链中广泛存在的问题，让人们能及时发现并解决问题，提高食品安全质量[16]。

## 1.3 本文研究内容与论文结构

本文使用网络爬虫技术爬取数据，对不同格式的数据进行处理，最后使用文本处理方法对文本数据进行研究。首先我通过爬虫获取所需的数据。然后我研究了对不同格式的数据进行处理的方法，能够快速提取出其传递的语义内容。最后我利用文本处理技术对提取到的内容进行处理，有效剔除不属于食品安全话题的垃圾信息，并对一定时间内的新闻报道数据进行聚类，直观快速地反映该段时间内食品安全的热点话题。

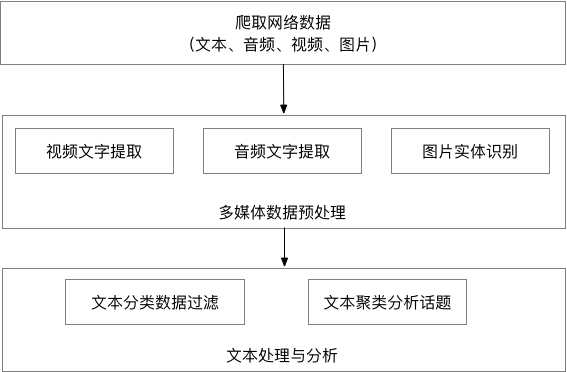


图 1.1 整体研究框架

本文的贡献主要有：

1.打破了传统对于食品安全领域研究仅限于文本形式网页的限制，能够对文本、音频、视频等多种格式的数据进行处理。

2.针对爬取到的数据中包含一些与食品安全话题无关的垃圾数据的问题，使用文本分类方法来进行建模从而判别文本的话题，能够有效过滤与食品安全话题无关的数据。

3.使用文本聚类技术来对食品安全新闻报道进行聚类，快速发现热点话题。

本文结构如下：

第一章是论文的绪论部分。主要介绍了论文的背景和研究意义，对食品安全领域研究的国内外现状进行了分析。

第二章介绍了本论文中运用到的一些相关技术，包括网络爬虫技术、多媒体数据的处理以及文本处理技术。

第三章主要描述了论文中利用爬虫技术在网络平台上爬取数据以及对音视频、图片数据的预处理过程。

第四章介绍了论文中使用文本分类技术来对数据进行筛选和过滤的方法。

第五章介绍了论文采用文本聚类的方式来对话题进行提取的过程，同时对采用LDA主题模型进行话题提取进行了讨论。

## 1.4 本章小结

本章主要介绍了本文的研究背景和意义，对国内外计算机领域学者对于食品安全话题的研究进行了总结，最后介绍了本文对于食品安全网络公开数据采集技术的研究内容，给出了本文的组织框架。在下一章节中我将对本文中使用的相关技术进行介绍。

# 第二章 相关技术

## 2.1 网络爬虫技术

网络爬虫技术是用来快速获取互联网中大量数据的一种方法，简单来说它就是从网络上下载信息存储到本地数据库的技术。不同于普通手动下载网页内容，网络爬虫是一种通过自动化方式来浏览互联网中的网页，检索web页面，然后下载信息的技术[17]。网络爬虫中，首先有一组初始url链接（种子url），通过访问这些链接，提取页面中的新链接存储进链接队列。根据链接队列，不断获取到新的页面下载所需内容，直到满足一定的终止条件。由于网络爬虫技术简单易学，可以快速方便地从网络上获取大量数据，因此它已经被各个领域广泛使用。

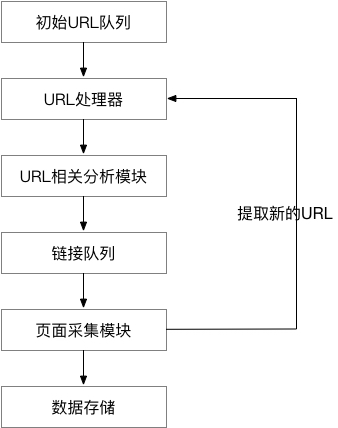


图 2.1 网络爬虫基本框架

### 2.1.1 爬虫分类

网络爬虫除了最普通的通用爬虫外，根据不同的需求衍生出了不同种类的爬虫，常见的有聚焦爬虫、增量爬虫、以及分布式爬虫等[18]。

聚焦爬虫是针对确定主题的爬取。当我们需要获取的数据只是某一特定领域的数据时，聚焦爬虫则能够满足我们的需求。它根据需求爬取指定的网页，在获取到网页后，不直接下载网页内容，而是先对网页内容进行研究分析，过滤一些无关主题的网页，再对保留下来的url链接进行爬取[19]。由于针对性更强，聚焦爬虫消耗的网络资源和爬虫时间远远小于普通爬虫，因此，在对数据有明确的主题需求时，人们通常采用聚焦爬虫。

而增量爬虫则是为了应对更新速度较快的网站而设计的一种爬虫方法。实际上，互联网中许多网站都是实时更新的，尤其是一些时效性较强的网站，如果按照传统的方式来爬取，很可能会丢失最新内容。当我们需要时效性很强的数据时，增量爬虫能通过监测网站的更新情况，顺利爬取到网站变化后的数据。

分布式爬虫主要针对一些商用的软件或程序，例如搜索引擎等。面对大量需要抓取的网页，如果采用传统的爬虫，所需要的时间消耗非常大，这会严重影响工作效率，因此人们开发出一种分布式爬虫，通过多进程来从网络中抓取并下载页面，从而在相对较短的时间内完成对海量网页的抓取。

### 2.1.2 爬虫工具

Python是近年来兴起的一种编程语言，由于它带有成熟的库和工具，我们通常采用Python来进行爬虫。在Python中已经有成熟的爬虫框架可供使用，也有方便的工具能够对网页进行解析，下面我将介绍几种常用的爬虫工具。

#### 2.1.2.1 Scrapy框架

最初Scrapy框架常被用于数据挖掘、数据监测和自动化测试等领域，后来渐渐被用于网络爬虫领域。由于使用了Twisted异步网络框架，使用 Scrapy来进行下载时速度有了很大提升。另外，Scrapy框架还包含了多种接口，能满足爬虫的各种需求。

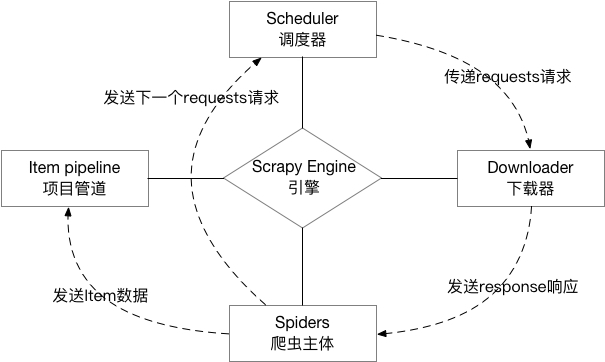


图 2.2 Scrapy框架

Scrapy框架主要是由五个部分构成，包括Engine（引擎）、Scheduler（调度器）、Downloader（下载器）、Spiders（爬虫主体）和Pipeline（项目管道）。Scrapy框架的核心是引擎，它负责处理系统中的数据流，可以用来触发事物。首先，调度器接收引擎发送给它的requests请求，压入调度器中的队列，当引擎再次发送一个请求给它的时候，它将从队列中返回一个请求传递给下载器。而下载器接收到引擎发送给它的请求之后，将会获取到response响应，并将响应发送给爬虫主体Spiders，爬虫主体在这个步骤中处理响应，并分析数据，从网页中提取数据Item，并将数据发送给项目管道，并发送下一个requests请求给引擎，从而进行下一次爬取。项目管道将会处理接收到的数据，进行数据处理或数据存储的工作。

#### 2.1.2.2 Selenium

Selenium是一种用于爬取反爬机制较严格的网站的工具。由于爬虫的人越来越多，被爬取的网站因为访问次数增多承受着巨大的压力，因此部分网站为了对站点性能进行保护会采取一些反爬虫措施。根据不同的反爬措施，人们提出了不同的绕过反爬虫机制的方法。其中Selenium就是爬取一些使用动态页面来进行反爬虫的网页的工具。

最初，Selenium是用于Web应用程序测试的开源自动化测试工具。由于它能通过模拟用户点击浏览器的操作来帮助开发者爬取到反爬机制非常严格的网站，人们在爬虫中也常使用到它。在使用Selenium的过程中，我们不考虑具体的请求和响应过程，只需要模拟人浏览网页然后获取数据的过程来下载数据。Selenium支持的浏览器包括IE、Mozilla Firefox、Safari、Google Chrome等。由于这种方法是模拟用户浏览网页来爬虫，因此时间消耗就会增多，一般只有在传统方式无效的情况下才采用这种方法来完成爬虫工作。

## 2.2 非结构化数据处理

由于网络中的数据多为非结构化数据，因此对非结构化数据进行处理十分重要。下面介绍几种处理非结构化数据的方法。

### 2.2.1 视频文字提取

视频是一种典型的非结构化数据，我们无法利用传统数据库对其进行存储和检索，并且视频内容往往只有通过播放才能被人们了解，这会使数据分析的时间变长，效率变低。因此便出现了检测视频中文字的技术。2013年，Zhang等[20]对文字检测的常用方法进行了分类和评估。此外，Yang等[21]提出一种识别幻灯片中文本的方法，可以将其运用到讲座视频的检索领域。针对中文文字的特殊性，Bai等[22]提出了一种基于灰度的方法来进行文字识别，他们通过灰度像素的处理直接从视频图像中提取出中文文字。

由于视频是动态的、连续播放的，直接从视频中提取文字具有很大难度。因此我们通常先将视频转换为单帧的图片，再利用图像文字识别技术对其中的文字进行提取。图像文字识别的过程包含文字区域检测和定位，从图像中分割文字，对文字进行识别这几个步骤。近年来，深度学习技术发展迅速，通过建立神经网络，深度学习能够模仿人脑对数据进行处理。在图像文字识别领域，深度学习相较于传统的文字识别技术效果更好。目前人们已经研究出较成熟的图片文字识别工具，例如百度文字识别、腾讯文字识别等。

### 2.2.2 音频文字提取

音频文字提取技术也可以称为语音识别。我们可以通过许多方法来进行语音识别，常见的有基于语音学和声学的方法，利用模板进行匹配的方法，还有利用神经网络来进行语音识别的方法。

要进行语音识别首先要了解音频的常见属性，音频包含采样率、通道数、位数、比特率等常见属性。其中采样频率就是每秒钟取得声音样本的次数，这个次数越高，音频的声音质量就更好。位数就是采样值，该属性将采样样本的幅度量化，位数越高，采样就越精确。而音频中的比特率就是指音频每秒的传输速率。在模板匹配的方法中，进行语音识别首先需要进行特征提取，也就是分析处理语音信号，去除与语音识别无关的信息，再对语音信号进行压缩。然后对模板进行训练，构造声学模型，再利用模型进行分类和判决。

目前的语音识别工具有百度语音识别、讯飞语音识别、腾讯云语音识别等。这些工具已经较为成熟，可以满足日常语音识别的需求。

## 2.3 文本处理技术

文本数据是最典型的、存在最多的非结构化数据。如何对文本数据进行分析和处理是我们面对的新问题。下面我将通过数据预处理、文本表示和文本分类与聚类这几个方面来描述文本处理的相关技术。

### 2.3.1 文本预处理

计算机是无法识别自然语言的，要想利用计算机对文本进行处理我们首先要进行数据预处理的工作，通过文本预处理中的分词和去停用词，可以将文本分割为独立的单词，便于后续向量化表示文本。分词就是将文本分割为一个个独立的单词，中文与英文不同，每个词语之间没有空格，因此在分词的时候，我们不能简单根据空格和标点符号来分词。

分词大致有以下几种方法，利用字符串匹配、基于理解和使用统计方法来分词。而中文的分词又分为基于词典、基于理解还有基于统计的机器学习算法这几种。目前已经有很多开源中文分词器可供使用，论文[23]中对Ansj、httpcws、IKAnalyzer、imdict-chinese-analyzer、mmseg4j、jieba、Paoding,盘古分词这8个开源分词器进行了研究和测评，由于Paoding分词在准确率、速度方面表现优秀，Jieba分词在Python中使用方便，这两种分词器使用的最为广泛。

通过分词操作，我们将文本分为了独立的单词，然而这些单词中存在许多我们不需要的词语。这些词语对文本的含义几乎没有影响，它们被称为停用词，由于这些词语的存在，会增大我们后续操作的时间消耗，也可能会对实验结果造成影响。因此我们利用去停用词这一步骤来过滤这些词语，保证单词集合的纯洁性。常见的停用词被分为两类，第一类是在所有文本中出现频率都很高的词例如“是”、“的”、“地”等词语，另一类是常见的词语例如“当然”、“并且”等。我们可以通过构建停用词表来过滤这些词语。

### 2.3.2 文本表示

文本表示就是将文本转化数学表达，便于计算机理解。

#### 2.3.2.1 特征选择和提取

由于将文本分割为单个单词之后，文本的特征空间维度很高，在利用机器学习算法进行处理时，运算复杂度和时间复杂度都会变大。想要将特征空间的维度降低就需要使用特征选择或者特征提取。特征选择是从原始的数据集中选择子集，原始特征数据集在这个过程中并没有改变。在文本处理领域，特征选择就是去除文本中一些无意义的词，包括出现次数很少或出现次数很多的词，选出一部分重要且具有代表性的词语来代表原始文本集，从而可以降低维度。文献[24]中提到，特征选择主要包含三类方法，分别是包装器、过滤和嵌入式方法。常见方法有文档频率DF(Document Frequency)、信息增益IG(Information Gain)、基尼指数GI(Gini Index)、卡方统计CHI(Chi square statistics)、最佳词语BT(Best Term)、模糊性度量AM(Ambiguity Measure)、特征选择器DFS(Distinguishing Feature Selector)。而特征提取则是从现有的文本特征空间中重构维度较小的新特征，原始的特征空间在这个过程中会产生变化。特征提取主要包含主成分分析PCA(Principal Component Analysis)、潜在语义分析LSI(Latent Semantic Indexing)和一些其他聚类方法。

#### 2.3.2.2 文本表示模型

文本表示就是把字符串转变为计算机能够理解的数学表达形式的过程。常用的表示模型有三类：向量空间模型VSM（Vector Space Model）、主题模型、神经网络模型。

向量空间模型（Bag of words），它将文本表示为向量。向量空间模型忽略文本中单词出现的顺序，只将文本看作若干个单词的集合。在构建向量空间模型时，每篇文本以单词为单位分割，每个单词都用一个向量来表示，从而每篇文档就被表示为了一个长向量。单词在文本中的重要性决定了该维向量的大小。

主题模型是从概率生成模型的角度来表示文本。它采用非监督学习对文本隐含的主题建模，能够从大量的数据中找出文字间的主题信息。这种模型可以自动搜寻文字里的主题。具有代表性的主题模型有两种，pLSA（Probabilistic Latent Semantic Analysis）和LDA（Latent Dirichlet Allocation）。

神经网络模型是利用神经网络表示文本。通常我们将它分为三类，基于词向量合成的模型、基于RNN/CNN的模型以及基于注意力机制的模型。其中基于词向量合成的模型被运用的最为广泛，具有代表性的有Word2vec模型。

### 2.3.3 文本分类

文本分类是利用有监督的机器学习分类算法识别出文本的类别，在使用预先标注好类别的训练集训练分类器之后，分类器能为每篇待分类文本分配一个对应的类别。在垃圾信息过滤、新闻分类、标注词语的词性等方面都运用到了文本分类技术。常用的文本分类器算法有朴素贝叶斯分类器、KNN算法、决策树算法、SVM算法等等。

朴素贝叶斯算法是一种有监督学习分类算法，它假设每个特征值独立，理论基础就是一个贝叶斯公式。公式中表示类别，X表示特征。需要算出在待分类项出现的情况下每个类的概率，然后选出概率最大的那个类别作为待分类项所属的类别。设是待分类的项集，为特征属性。是已存在的类别集合。

（2.1）

SVM（Support Vector Machine）算法是解决二分类问题的常用算法，在SVM模型中，我们需要在特征空间里确定一个分离超平面，这个平面能够把训练集中的文本划分为两类。当类别边界沿垂直于该超平面的方向距离越大，那么分类的结果越准确。

决策树算法是通过建立树的结构来表现决策规则，展示分类结果。在决策树中，根节点代表对最终分类结果贡献最大的属性，叶子结点代表了最终的分类结果。

KNN算法也是一种常见的分类算法，对应给定的测试文档，我们在训练集中搜索，找到与这个测试文档最相似的K个文档，查看这K个文档的类别，从而判定测试文档的类别。

### 2.3.4 文本聚类

文本聚类是一种无监督的机器学习算法。我们不需要提前为训练集标注类别，算法直接比较一组文档的相似度，相似度高的分为同一类。类别是在聚类的过程中自动生成的。文本聚类包括划分法、层次法、基于密度的方法、基于网格的方法还有基于模型的方法。

其中划分法中典型的有K-means算法，主要思想就是通过分裂，将数据集分为多个组，每个分组就代表了一种类别。层次法中典型的有BRICH算法，不同于划分法中的分裂，层次法是先将每个文档都看做一个原子，对这些原子一层一层地进行聚类，直到满足一定的条件终止。基于密度的方法中，DBSCAN算法运用较为广泛，该算法的思想是把密度较高的且相近的点连成一片，从而生成簇也就是不同的类别。基于网格的方法要形成网格结构，聚类都在网格上进行。基于模型的方法则是为每个聚类假设一个模型，通过寻找满足该模型的数据集来进行聚类。

## 2.4 本章小结

本章主要介绍了本文中涉及到的相关技术。首先我介绍了网络爬虫技术的概念和类别以及一些常用的爬虫工具。然后介绍了对非结构化数据的处理，包含视频文字提取和音频文字提取两个方面。最后介绍了文本处理技术：文本分类技术和文本聚类技术。在下一章，我将介绍本文中对于食品安全网络公开数据的采集，从使用爬虫获取数据，利用MongDB存储数据和对多媒体数据的处理三个方面来展开介绍。

# 第三章 数据采集

本章我将从数据爬取、数据存储和对多媒体数据的处理这三个方面来介绍数据采集工作。本论文爬虫选用的主要数据源一共有7个。其中包含文本类型网页：中国食品安全网、食品科技网、国家食品药品监督管理总局平台；视频类型网站：梨视频、哔哩哔哩网站；音频类型网站：喜马拉雅FM、蜻蜓FM。

## 3.1 基于Scrapy的爬虫

由于Scrapy框架对于爬虫者来说十分友好，能大大提高我们的工作效率，因此本文中也采用了这种方式来对数据进行爬取。下面我将以爬取梨视频为例来介绍利用Scrapy框架来进行爬虫的步骤。

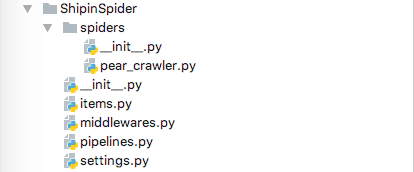


图 3.1 梨视频爬虫框架

首先，在item.py中建一个定义所需数据属性的类。然后在spiders文件夹中新建爬虫主体文件为pear\_crawler.py，在这里完成提取网页信息并获取新的url的工作。在pipelines.py文件中，对spider收集好的数据item项进行存储的处理。在settings.py文件中启动pipeline管道，如果需要将数据存入数据库则可以在这里进行配置数据库的操作。当我们完成上述所有操作后，就可以直接运行scrapy命令来对目标网页进行爬取。

## 3.2 MongoDB存储数据

由于MongoDB是介于关系数据库和非关系数据库之间的一种数据库，可以便于我们存储比较复杂的数据类型，并且它的基于分布式文件存储方式能够满足可扩展并且性能高的数据存储需求。因此本文中的所有数据都利用MongoDB来进行存储。在MongoDB中，用集合来代替关系数据库中的表，每个集合在数据库中有唯一的标识名，可以存储无限的无需定义任何模式的数据。

MongoDB中，在进行数据存储时，不需要提前建立数据库和集合，在进行数据插入操作的同时，若没有该数据库或该集合，MongoDB会自动创建。被插入的数据称为文档，也就是一组键值对，我们不需要为文档设置相同的字段，甚至不需要为相同的字段设置相同的数据类型。这让数据存储工作变得非常简单高效。

## 3.3 多媒体数据处理

本文中需要处理的多媒体数据包括视频、音频以及图片。其中视频、音频数据都属于流式数据，我们很难直接获取其中的内容信息，因此本文提出通过提取音视频中的文字来代表其传递的内容。而图片作为一种最为广泛传播的数据，也可以利用图像识别技术来对其进行识别，提取其包含的内容。

### 3.3.1 视频文字提取

近年来，短视频越来越流行，由于时长短、内容精悍，人们往往乐于传播分享这类信息。本文以梨视频和哔哩哔哩作为数据源，以“食品安全”作为关键词对视频进行爬取，将爬取到的视频信息包括title（视频标题），pubdate（发布时间），sourceurl（视频地址源链接）存入MongDB数据库中。视频标题一般字数较少，虽然能简要概括视频内容，但对我们了解视频的本质内容没有什么帮助，而视频中的字幕却能很好地反映视频的语义信息。因此只需要将视频中的字幕提取出来就可以快速了解视频内容。第二章中介绍了关于视频文字提取的办法，但是由于视频图像背景复杂、颜色丰富，传统文字提取算法很难达到很好的效果，并且采用传统算法来提取文字会导致程序消耗时间增多。因此本文采用了目前发展较为成熟的百度AI通用文字识别来对视频进行文字提取。百度AI通用文字识别能够识别图片中各种语言的文字，能够准确识别出所有常用字和大部分的生僻字，并且提供了高精度和包含了位置信息的版本可供使用。其在拍照截图、内容审核与监管、视频内容分析和纸质文档电子化这些方面都有所应用。

视频可以看做由多个帧组成的，在提取视频中的文字时，我们首先把视频分割为单帧的图片。Python中有imageio库提供简单接口来对各种图像数据进行读取和写入操作，也有ffmpeg工具来对音视频进行处理。本文中，我采用Python所带的imageio库来读取视频，利用ffmpeg来解码音视频文件。通过对视频数据的观察我们发现，每个视频中包含的帧数有很多，一个一分三十秒的视频中包含的帧数就会有两千帧之多，如果将每一帧都截取出来，程序运行时间会大大增加。另外，并非每个视频帧中都包含字幕，通过观察我们发现，视频中字幕的交替出现时间间隔通常为2秒，我们每隔2秒保存一次视频帧图片，能够保证视频中包含字幕的帧都被截取，并且能尽量减少重复的字幕出现。下面我将对视频文字提取的步骤进行详细介绍。

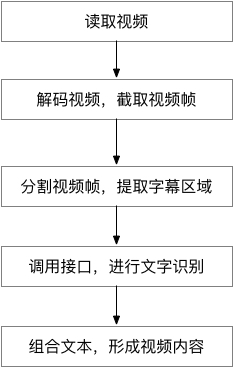


图 3.2 视频文字提取过程

首先，读取视频流，通过设置帧数间隔，遍历视频流，从视频流中截取指定帧，然后将指定帧保存下来。在这里，如果帧数间隔过小，就会保留许多重复的帧，如果帧数间隔过大，则可能会丢失一些重要信息。通过对帧数间隔的大小进行测试，我设定帧数间隔为50，每50帧保存一次。

在将视频分割为图片之后，需要提取出图片中的字幕部分。通过观察，本文中所使用的数据源梨视频和哔哩哔哩平台上的字幕都位于视频下方部分。因此在截取视频帧时，我们设置截取区域为图片下方部分，这样在后续对图片进行识别时就不会识别到无关的对视频语义内容影响很小的类似于“梨视频”、“哔哩哔哩”等对视频内容无用的文字。同样地，当视频字幕位于视频上方或者其它固定的位置，我们也可以通过截取视频帧的固定区域，来减少对不相关文字的识别。然而，并非所有的视频字幕都位于视频中的固定区域，在一些视频中，能表现其内容的文字在每一帧中出现的位置可能会有很大变化。这时，我们可以先识别整个图片中包含的文字，再对识别到的内容进行预处理，过滤掉那些对视频内容影响很小的文字。

本论文采用百度AI开放平台的SDK客户端来进行文字提取工作，在识别单张图片中的文字之后，将所有图片中的文字组合起来，形成视频的内容文本，该文本能反映出视频的主要语义信息。除了文字，有一些符号也被识别进内容中。除此之外，一些句子因为在视频中停留的时间过长，因此被重复识别，这样的字幕文字内容还经过后续处理才能使用。

### 3.3.2 音频识别

本文中的音频通过对喜马拉雅FM和蜻蜓FM平台进行爬虫得到，这两个平台都是国内较知名的音频分享型平台，其包含的内容涵盖了各个方面，也不乏关于食品安全的音频。首先 “食品安全”搜索结果中爬取下音频，在MongoDB数据库中，存储每个音频的title（视频标题），pubdate（发布时间）， url（音频链接）sourceurl（音频地址源链接）。然后我们需要从音频中识别出其包含的文字来代表音频的内容。我们将音频下载到本地，以便后续的音频识别操作。目前语音识别技术发展较为完善，效果较好的语音识别工具有百度语音识别、科大讯飞语音识别、腾讯语音识别等。本论文选用百度语音识别来对音频进行识别。由于百度语音识别技术API对音频格式要求严格，它支持的语音格式只有pcm、wav、amr三种格式，并且规定原始pcm的录音参数必须符合 8k/16k 采样率、16bit 位深、单声道，支持的语音时长上限为60s。而我们获取到的音频为m4a格式，时长各异，音频参数也不符合要求，因此在进行音频识别前，我们首先需要对原始的音频进行转换。

上一小节中我们了解到ffmpeg是一个开源的用来处理音视频流的工具。这一小节，我使用ffmpeg命令完成对音频格式的批量转化。首先将m4a为后缀的音频文件转换为wav格式的音频。再对音频进行分割处理，将它分割为多个时长为60s的短音频，然后利用ffmpeg命令将这些音频片段的参数进行变换使其符合百度云音频识别的参数要求。完成了对音频的格式转换之后，我们就可以识别音频中的文字内容。

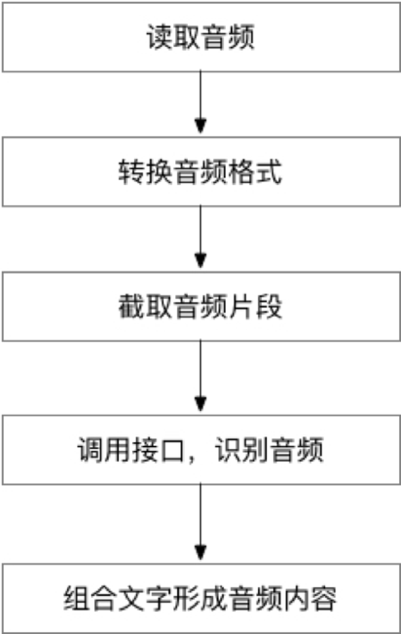


图 3.3 音频识别过程

### 3.3.3 图片实体识别

在一些社交网站或者APP中，为了简单方便，人们传播信息时可能会直接传播图片，而不对其进行文字描述。由于人们对食品安全领域的关注度越来越高，因此食品安全方面的信息传播也逐渐增多。其中，餐厅后厨往往是人们重点关注的话题，后厨是否干净整洁可以从侧面反映出该餐厅的食品是否安全卫生。在一些不符合卫生标准的餐厅后厨中，经常会进入一些对食品卫生造成影响的动物，例如老鼠、猫、狗等。使用图像识别技术我们可以识别出图片中的动物，这一技术可以运用到监管餐饮行业后厨中。另外，识别图片中的食物也是一个值得研究的领域。在一些食品安全问题事件出现时，网络上会出现大量关于该食品的图片，如果我们能快速识别图片中的食物种类或者名称，就能及时了解当前食品安全事件，帮助人们了解当前事态发展并采取相应的应对措施。本文中我采用百度AI图像识别接口完成了对图片实体的识别，识别对象主要包含动物和食物。

我首先下载了百度AI平台提供的图像识别SDK文件到本地，该接口不仅能识别通用物体，还能识别细粒度物品，例如菜品、车型、logo商标和动植物等。本文中主要调用了通用物体、菜品以及动物的识别接口。首先判断图片中的实体是否是动物和菜品，若属于这两类，则进行识别，将识别结果进行存储。若判断该图片既不属于动物也不属于菜品，则调用通用物体识别接口返回识别结果进行存储。本文中，我主要存储了图片实体的名称，该实体在百度百科中的介绍链接以及对该实体的文字描述信息。

我们可以利用识别图片中的动物来判断餐厅后厨是否符合食品安全规范，帮助人们更高效地对餐厅后厨的卫生情况进行监管，保障人们的食品安全。另外，我们可以将识别图片中的食物利用到食品安全监管中来，当网络上大量出现关于某食品的图片时，我们能快速获取关于该食物的饮食禁忌或者食品安全标准，提醒人们注意食品安全。

## 3.4 实验与分析

在视频文字提取模块，我利用3.3.1节描述的方法进行了实验，提取了一共726条视频的内容。通过实验我们发现，使用本文中的方法能够在消耗时间比原视频时长较少的情况下，识别出视频传递的主要内容。图3.4为对一个时长为1分22秒的视频提取文字的结果，消耗时间大约为49秒。与视频中原本的内容进行对比，发现提取出的文字能够反映视频的内容。对于识别到的字幕，如何判断识别是否准确成为了新的问题。我将从识别的准确率，漏检率这两个方面，通过人工检查的方式，将识别到的文字和原视频进行比对。由于在识别单张图片时，不能保证每个字都被识别准确，因此当识别准确的字数占句子总字数的80%时，我们就认为该句子被准确识别。在这里，我将准确率定义为识别正确的句子占原视频中句子总数的比例；漏检率代表没有识别到的句子占原视频中句子总数的比例。在这里，我测试了从梨视频爬取到的5个不同时长的视频，包含“食品快检车开进超市”、“黑外卖脏乱差”、“老鼠钻食品柜啃鸡翅”、“厕所旁熏腊肉”、“饿了么开直播”。首先人工观看视频，将视频中的字幕记录下来，统计其包含的句子数目，然后将原视频中的每个句子与识别到的结果进行逐句对比，统计识别正确和被漏识别的句子数目。通过将识别到的字幕与原视频进行对比，我们得到本文视频字幕识别准确率为74%左右，漏检率为22%左右。也就是说，我们在不播放视频的情况下就能了解其传递的基本语义信息。这将会在实际应用中大大提高效率，节省人力。

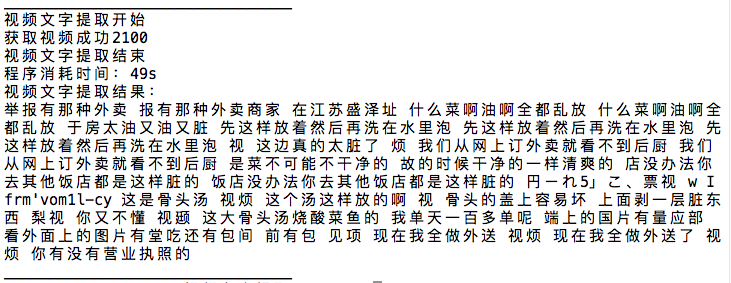


图 3.4 某视频的文字提取结果

在音频识别模块，我利用3.3.2节的方法进行了实验。通过对不同时长音频进行测试，我们发现，利用本程序识别一个1分钟的音频只需10秒；识别5分钟左右的音频只需要65秒；识别10分钟左右的音频只需要150秒。通过程序识别音频内容所需要的时间远远小于直接播放音频消耗的时间，能大大提高工作效率。图3.5为一个1分14秒的音频提取出的文字结果，程序消耗时间为18秒。我们发现识别结果虽然不是百分之百准确，但也能基本反映音频所传达的语义信息。在不播放音频的情况下，我们就能在较短时间内了解音频传递的信息，这大大提高了人们对音频文件的处理效率。同样地，我们将识别到的文字与原音频内容进行对比来评估音频识别的效果。音频与视频不同，视频在分帧的过程中会存在截取重复帧或者漏截取部分包含字幕的帧，因此会出现一些句子在识别结果中反复出现，或者一些句子在识别结果中没有出现的情况。而音频是直接按间隔60s截取片段，因此不存在重复或者漏掉的问题。所以我们在评估音频识别时，只需考虑识别到的内容是否与原音频内容相符，本文中我将利用准确率来评估音频识别的效果。准确率代表音频中识别正确的词数占音频总词数的比例。在此，我测试了5个从蜻蜓FM爬取到的音频，包含“食品追溯的重要性”、“永辉超市食品安全不合格”、“食品安全法规”、“食品安全的底线不能去海底捞”、“崔永元关注食品安全”。通过人工播放音频，记录下原音频的文字内容，统计音频中包含的词语数目，再将识别到的文字与原音频的内容进行比对，比较每个词语是否被正确识别，统计被正确识别的词语数目。通过将音频识别到的内容与原音频进行对比，在原音频为标准且清晰的普通话的情况下，我们得到音频识别的准确率大约为94%。音频识别的准确度与原音频的质量有很大关系，如果原音频中发音不标准，吐词不清晰，那么识别的准确率就会相应降低。

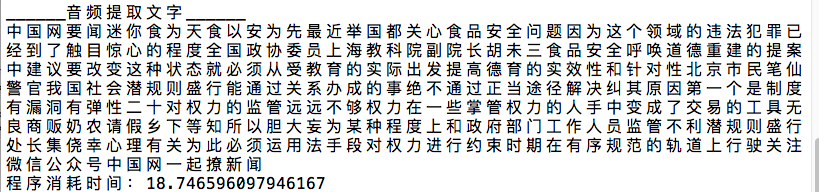


图 3.5 某音频的识别结果

最后，在图片实体识别中，本文主要是针对动物、食物进行识别。借助百度AI平台，我们可以识别五万种不同的菜品种类，能够识别数千种不同类别的动物。对多张图片的测试结果表明，只要图片清晰且物体在图片中占比正常，程序就能对其进行准确识别，并能显示出对于该图片的详细描述和解释。在很多情况下，人工对图片进行辨别的准确率并不高，例如需要辨别的图片是不常见的动物或者菜品时，人工就很难识别。另外，人工识别图片会消耗大量的人力和时间，这在图片数量庞大时是不可行的。本文中的方法能够让计算机快速准确返回被识别图片的内容，从而能在更多场景下运用。我对不同类别的菜品进行了测试，图3.6为学校食堂中常见的三种菜品，图3.7为对这几种菜品的识别结果。可以发现，该程序能够准确的返回菜品信息。



图 3.6 学校食堂常见菜品

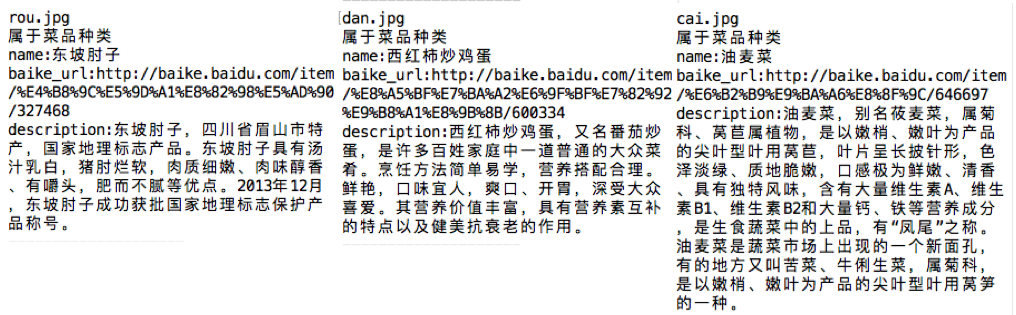


图 3.7 对菜品的识别结果

同样地，我对不同类别的动物进行了测试，发现该程序对动物的识别也十分准确，图3.8为日常生活中我们可能会在厨房看到的三种动物，图3.9为对这三类动物的识别结果。可以看到，程序识别图片往往比人工识别更加精确快速，我们也可以得到更多关于该图片的信息。



图 3.8 常见动物

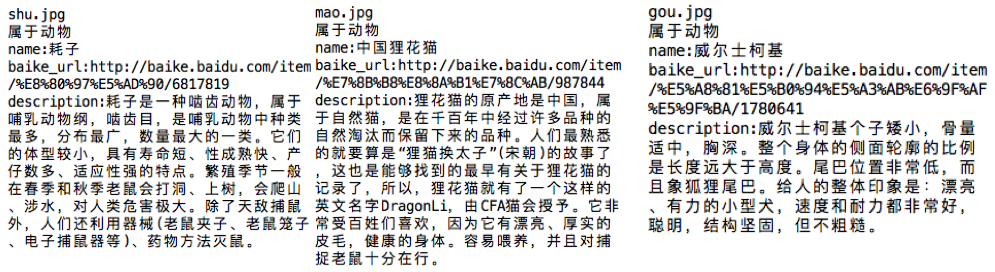


图 3.9 对动物的识别结果

## 3.5 本章小结

本章从爬虫、数据存储和多媒体数据处理三个方面介绍了数据采集部分的工作。由于爬虫得到的数据格式多样，除了文本还有音视频以及图片，本章讨论了对音视频、图片这三种类型的处理。从音视频中提取文字使我们不用通过播放便可以了解音视频传递的语义信息。而图像识别可以快速判断人们传递的图片是否与食品安全相关。在下一章，我将介绍本文中使用文本分类来进行数据过滤的方法。

# 第四章 利用文本分类进行数据过滤

第二章中，我们介绍了文本分类的相关技术。文本分类能通过训练好的分类器来判别输入文章的类别，这可以用来预测文章的主题。本文根据文本分类的特点来对数据进行过滤。虽然我们在从网络上爬虫的时候预定义了爬虫的关键词，但这仍无法保证我们得到的数据都是与食品安全话题相关的内容，本章中，我将详细描述本文利用文本分类技术对数据进行筛选，只保留与食品安全相关的文章的过程。

## 4.1 文本预处理

在第三章的数据采集过程，我们从食品安全网和CCTV网共爬取了八百余条关于食品安全的新闻报道，将来自于这两个数据源的报道组合形成与食品安全话题相关的数据。另外，我爬取了CCTV网国际、国内、军事、科技、社会、法律、娱乐、经济这些领域的一千余条数据。由于计算机无法直接处理文本内容，我们需要将其转化为计算机能够理解的数学表达。

通常，文本中包含了词语、符号、数字、英文单词等。在本论文中我们主要研究中文文本，因此我首先利用正则表达式分离出文本中的中文文字，去除对文本语义影响不大的英文和符号。然后采用Python中的Jieba分词来进行分词操作。分词之后，文本就被分为了独立的单词，但是在这些单词中，存在许多对文本语义影响很小但出现次数却很多的词语，例如“为”、“对”、“等”、“的”等词语，这些词语虽然对文本语义贡献微小，但会对后续的分析结果造成严重的影响。目前人们总结了各种停用词表来供我们使用，在本文中我通过遍历百度停用词表，删除文本中的停用词。

## 4.2 基于文本分类的主题判别

本文通过文本分类来确定待预测文本的主题，也就是说本文中的文本分类实际上可以看做一个二分类问题。数据集中属于食品安全话题的为第一类，不属于食品安全话题的为第二类。当采用训练集数据训练好分类器之后，分类器可以对输入的文本进行判别，判断其是否为食品安全话题数据。

### 4.2.1 TextRank提取关键词

第二章中我们介绍了特征选择方法，可以选取对文本内容影响较大的单词来代表整篇文本，在本论文中我选择提取文本的关键词来代表文本内容。在自然语言处理领域，提取关键词是一个非常基础且常见的步骤。提取关键词的方法也有很多，常用的有根据TF-IDF来提取关键词和利用TextRank提取关键词。由于基于TF-IDF的方法只考虑了词的统计信息，而不考虑词之间的语义信息，具有一定的缺陷，因此在本文中我采用TextRank算法来对文本关键词进行提取。

TextRank算法是一种基于PageRank的关键词提取方法，它利用图的思想，将文本中的词语表示为无向有权图，在图中，每个节点表示一个词语，每个权重表示词语与词语之间的联系程度。提取前N个关键词的过程也就是从图中找出权重排名前N的节点。简单来说，在TextRank算法中，如果一个词语出现在很多词语之后，那么它的重要性就比较高。由于TextRank算法考虑到了词语与词语之间的关联性，不是单纯靠统计信息来确定关键词，因此这种方法相比其他方法更加科学。

### 4.2.2 文本向量化

通过上一节的文本预处理，我们已经将文本转化为了独立的词语，下一步则是将文本向量化为计算机能够理解并处理的数学表达。第二章中我们已经介绍过各种文本表示模型，其中向量空间模型的概念较为简单，实现起来也比较方便，因此本论文中采用了向量空间模型的方法，在其基础上进行了改进，从而将文本向量化。

在传统的向量空间模型中，训练集中的每个单词都作为一个维度。在这里我将用一个例子来说明，如果有一个词向量空间为（我，爱，冰淇淋，她，不爱），那么句子“我爱冰淇淋”则可以用向量表示为（1，1，1，0，0），相应的，句子“她不爱冰淇淋”可以用向量表示为（0，0，1，1，1）。此时，这些向量都是位于同一个向量空间，因此计算机可以对其进行各种运算。但是在实际情况中，我们的训练集是成百上千甚至上万的文本，每个文本中包含词语也不计其数。文本集中的词语数目巨大，如果被直接用来构建向量空间模型，那么会使得向量空间的维数变得很大，造成维度灾难，甚至会增加后续的计算难度和程序的运行时间。因此直接简单地利用这些数据来建立文本模型是不实际的，本文从每篇文本中选出一定数量的关键词来代表整个文本，这些关键词就是该文本的特征词，从而达到特征选择的目的，使向量空间的维度降低。

我们将数据集中的文本划分为两类，属于食品安全话题的为第一类（主题文本集），不属于食品安全话题的为第二类（垃圾文本集）。首先，我需要构建主题词典来方便后面构建统一的向量空间。在构建主题词典时，我先在主题文本集中利用TextRank算法提取每个文本的前top-K个关键词，再从这些关键词集合中选择出现次数最多的前top-N个词语，组合这些词语形成主题词典。这一步的目的是为了降低向量空间的维度，便于后续的计算，提高计算的速度。此时，我们得到了大小为top-N的主题词典，词典中每个词语对应一个权值，该权值就是词语在关键词集合中出现的次数。向量空间模型中，每个文本都将被表示为一个向量，在此，我将介绍对一篇文本进行向量化的过程。首先，我们利用TextRank算法提取该文本的前top-N个关键词，将其与主题词典中的词进行一一对比，当某关键词与主题词典中的词语一致，那么对该位置赋值，数值大小为主题词典中该词的权值。遍历一遍关键词组之后，我们就得到了每个位置上的权值，由此我们就得到了该文档的向量表示。对文本集中的所有文本都进行这样的操作之后，每篇文本都被转换为了向量。这一步实际上是对传统的向量空间模型的改进，用主题词典来代替传统的词向量空间，用关键词组来代替每篇文本，从而可以达到降维的目的。

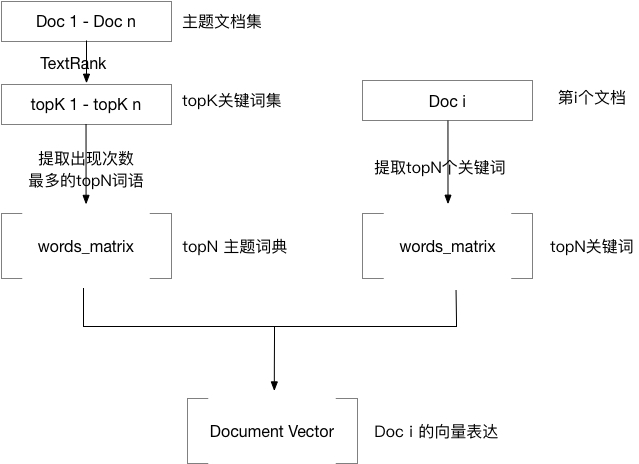


图 4.1 文档表示为向量的过程

### 4.2.3 构建分类器

在第二章中我们已经介绍了，文本分类有许多分类器可供我们使用。本文的目的是通过文本分类来确定待预测文本的主题，也就是说本文中的文本分类实际上可以看作是一个二分类问题。数据集中属于食品安全话题的为第一类，不属于食品安全话题的为第二类。SVM（Support Vector Machine支持向量机）分类算法是一种被广泛应用于二分类问题的算法，由于原理简单，使用方便，本论文中也将采用SVM算法和贝叶斯算法来进行文本分类，并对比这两种分类器在本文中的分类效果。

上一小节我们已经得到了每篇文档的向量表示，在构建分类器时，首先将数据集分为两类，一类为食品安全话题下的数据，一类为其他数据。数据集被分为两部分，一部分为训练集，一部分为测试集。将训练集中的文档向量输入我用Python编码实现的分类器中，调整参数训练模型，得到分类器。分类器的准确度和训练模型时设置的参数，训练集的大小，训练语料的准确度等方面都有关。

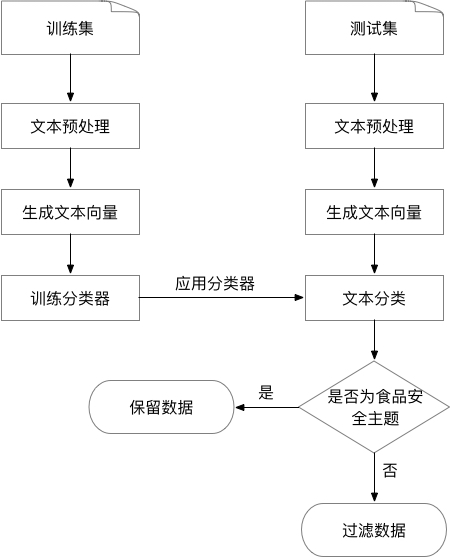


图 4.2 分类器过滤数据过程

## 4.3 实验与分析

在4.1节中已经介绍了文本预处理的基本方法，在此我对同一个文本进行分词和去停用词的结果进行展示，从中可以看出，一些对文本内容几乎没有影响的词例如“从”、“开始”等词被顺利剔除。

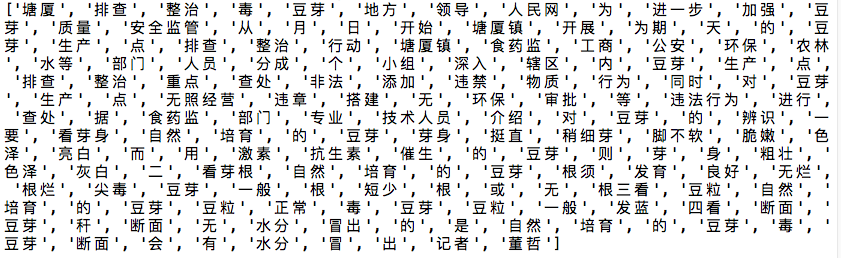


图 4.3 分词后的文本

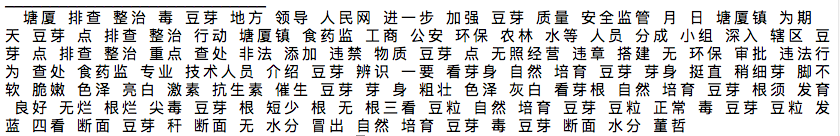


图 4.4 去停用词后的文本

由于食品安全类的新闻报道占据总新闻报道的比例未知，我们的问题是一种不均衡的二分类问题。为了规范，在本文中选择训练集数据时，我以1:1的比例从食品安全数据和非食品安全数据中各抽取800条作为数据集。首先，我将食品安全主题下的文档集都标记为1，非食品安全主题下的文档集都标记为0，再将数据分为训练集和测试集，其中训练集占80%，测试集占20%。我使用两种不同的分类器来对测试集进行测试。在文本分类中，我们可以通过混淆矩阵来观察分类器的预测效果。本文中，混淆矩阵中横坐标表示预测值，纵坐标表示实际值。图4.5中展示了朴素贝叶斯分类器和SVM分类器的混淆矩阵。在分类任务中，常见的评估分类器效果的指标有准确度（Accuracy）、精确度（Precision）、召回率（Recall）等。本文中，准确度表示被分类正确的样本占总样本的比例；精确度表示实际为食品安全话题的样本占被分类器判断为食品安全话题样本的比例；召回率表示被分类器判断为食品安全话题的样本占实际为食品安全话题样本的比例。表4.1中展示了两种分类器的性能指标。通过表4.1我们可以看出SVM分类器的准确率、精确度和召回率都要略高于朴素贝叶斯分类器。利用SVM分类器来判断文章是否属于食品安全话题时，召回率高达83.8%。

表 4.1 朴素贝叶斯分类器和SVM分类器评估指标

|  |  |  |
| --- | --- | --- |
|  | 朴素贝叶斯分类器 | SVM分类器 |
| 准确率Accuracy | 0.881 | 0.903 |
| 精确度Precision | 0.949 | 0.956 |
| 召回率Recall | 0.80 | 0.838 |

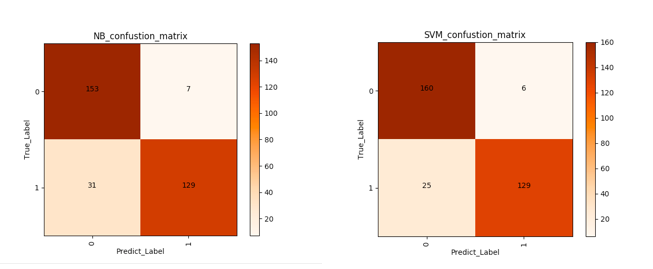


图 4.5 左图：朴素贝叶斯分类器 右图：SVM分类器

图4.6展示了利用SVM分类器来判断文本是否属于食品安全话题的部分实验结果。我从属于食品安全话题的数据和不属于食品安全话题的数据中各抽取20条，用本文中的SVM分类器进行主题判别，可以看出，有17个食品安全话题数据被正确判别，有3个食品安全话题数据被漏判，有4个非食品安全数据被误判。



图 4.6 SVM分类器测试结果

上述实验结果证明了本文提出的利用文本分类来进行主题判别是可行的，并且达到了较好的效果，其中SVM分类器的精确度达到了95.6%，召回率达到了83.8%。对一篇输入的文本，利用本文中的分类器能够较为快速准确地判断其是否与食品安全话题相关。当数据集十分庞大时，我们就可以利用本文中的分类器进行筛选和过滤，能够快速并且较为准确地删除一些与食品安全无关的数据。当然，这一工作也可以被拓展到其他领域，我们只需要改变训练数据集，就能轻松的将分类器用于其他领域的主题判别和数据过滤。当然，除了本文使用的朴素贝叶斯和SVM分类算法，还有很多其他分类算法可以使用，例如决策树、神经网络，K-近邻算法，以及深度学习的方法等。然而这些方法使用起来比朴素贝叶斯和SVM算法更加复杂，且本文使用的这两种方法已经能够对数据进行有效的分类，能够完成主题判别的任务，因此本文没有采用其他的分类算法进行实验。如果我们对分类器的性能和准确度有更高的要求，希望找到最好的方法，那么我们可以考虑采用其他更多不同的分类算法来进行实验，比较分类器的准确度、精确率、召回率以及时间消耗，从而选择更适用于本文主题判别的方法。

## 4.4 本章小结

本章介绍了利用文本分类进行主题判别的方法，首先对文本进行预处理，然后将每一篇文档表示为向量，构建了SVM和朴素贝叶斯两种分类器，对输入的测试数据进行了分类，实验结果表明，本文中的分类器能有效判别一篇文档是否属于食品安全话题。

# 第五章 话题提取

由于从互联网中采集到的数据数量庞大，且内容各异，我们通过第四章中介绍的数据筛选，只能过滤掉一些与食品安全无关的数据，保留下的数据依然很难进行直观分析。在自然语言处理领域，人们通常采用文本聚类技术对大型文本集中隐含的潜在主题进行挖掘，由于我们的数据集包含了各个领域，没有事先标记好类别，因此文本聚类这种不需要事先标注类别的方法就特别适用于这种情况。文本聚类的思想和普通聚类算法的思想一致，就是通过聚类，使不同的类之间的距离越大越好，而相同类之间的距离越小越好。

## 5.1 K-means文本聚类分析

K-means是一种典型的基于划分思想的聚类算法。下面我将介绍利用K-means来进行文本聚类的过程。

### 5.1.1 K-means算法

我们需要先确定聚类的数目，然后随机选定一些初始聚类中心，计算每个点和初始聚类中心的距离，将其归类，然后计算每个类的中心作为聚类中心，重复上述的操作，当聚类中心确定下来不再改变的时候，算法就终止。这里需要说明的是，K-means算法中通常采用的距离计算为对欧式距离进行计算。通过上述过程，我们可以使得同一类中的样本点的距离尽可能小，不同类之间的样本点的距离尽可能大。K-means算法的步骤如下：

1. 确定聚类的数目K。
2. 随机选择K个样本点作为聚类的初始中心。
3. 计算每个点与每个初始中心的欧氏距离，将其划分到距离最小的类中。
4. 计算每个类的中心作为下一次的聚类中心。
5. 重复步骤（2）（3）（4），当聚类中心不变时，算法终止。

### 5.1.2 利用K-means进行文本聚类

文本聚类又被分为了划分法、层次法、基于密度、基于网格、基于模型的方法。其中划分法中的K-means算法由于原理简单、便于上手并且时间复杂度较低，因此人们常常利用它来对文本进行聚类。本论文中也将采用K-means算法来对采集到的数据进行聚类分析从而提取出关键性的话题。

#### 5.1.2.1 TF-IDF向量化文本

文本聚类同文本分类一样，首先需要将文本转化为计算机能够理解的数学表达，也就是将文本向量化。在第四章中我们介绍了本文中选用的文本向量化方法，通过建立主题词典来对每一篇文章进行向量表达。而在文本聚类中，由于数据本身是无标记的，我们无法通过这种方式来向量化文本。因此本文中采用了基于统计的思想也就是TF-IDF来对文本进行表达。

TF-IDF（Term Frequency – Inverse Document Frequencey）也叫词频-逆文档频率。由于传统的统计词频方法具有严重的缺陷，某些词语虽然出现次数不多，但对文本语义影响重大，这些词语在传统方法中往往会被分配到很低的权重，造成后续分析结果不准确。因此人们提出TF-IDF的思想。

（5.1）

（5.2）

公式（5.1）中TF(x)代表词x的频率，IDF(x)代表词x的逆文档频率。公式（5.2）中N代表文章的总数目，N(x)表示了语料库中包含了词语x的文章数目。使用TF-IDF来计算词语的权重时，当某词在某篇文章中出现的次数越多，它的重要性就越高，而在语料库中出现的次数越多，它的重要性就越低。利用TF-IDF计算词语的权重克服了用词频来计算权重的缺点，能简单快速地对词语权重进行计算从而获得文本的向量表达。

本文通过计算TF-IDF矩阵得到特征矩阵，其中第一列(i,j)表示语料库中的第i篇文章，语料库中的第j个词语，第二列的数值代表了该词语所占的权重。通过上面得到的TF-IDF值，每篇文章都可以被表示为矩阵，其中每个词语的权重就是TF-IDF值。

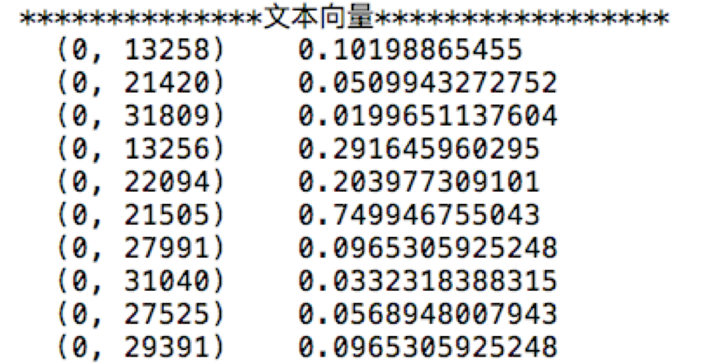


图 5.1计算的TF-IDF值

#### 5.1.2.2 文本聚类

通过上一小节的计算，我们已经将每篇文本都转化为了相应的向量表达。通过使用K-means算法，将文章向量作为输入，来对其进行聚类。由于K-means聚类需要预先确定聚类的数目，由于本文只研究食品安全领域的文本内容，因此如果设置的聚类数目太小，那么聚类效果就没有意义，而聚类数目如果太大，也会造成话题数目过多，违背了聚类分析话题的初衷。通过综合考虑以及对比实验，本文中我将聚类数目设置为15，为了体现聚类结果，在得到聚类每篇文章所属的簇之后，我对同一簇的文本进行了关键词提取，目的在于发现该簇表达的话题信息。

通过文本聚类，我们可以快速高效地了解一段时间内食品安全的热点话题，可以帮助相关及时了解当前存在的问题，并及时采取措施控制局面，将食品安全问题带给人们的影响降到最低。

## 5.2 LDA分析话题

由于传统的聚类分析大多是基于统计学的方法来对文本进行分析，而忽略了词语本身的含义。因此人们提出使用主题模型来对话题进行提取，一种典型的主题模型就是LDA（Latent Dirichlet Allocation）。文献[10]为从互联网新闻中提取信息并分析其隐含的主题，提出了一种把LDA主题模型和主题评价方法结合的方法。文献[11]试图通过LDA模型来挖掘图书馆中电子文献的研究主题和趋势，文中提到可以根据困惑度、经验似然等方法来对主题数目K进行指定。文献[12]中主要研究了类似于学术摘要这样的短文本的聚类，将Word2Vec和LDA模型进行了结合。而文献[13]也利用LDA模型对从网络上爬取到的新闻数据进行了聚类分析。

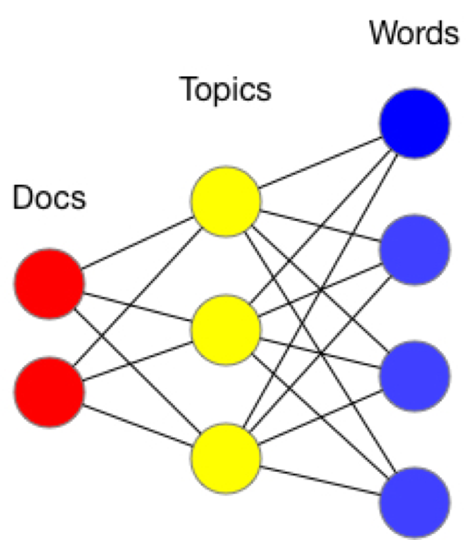


图 5.2 LDA模型

LDA（Latent Dirichlet Allocation）是主题模型中的典型，它是一种隐含狄利克雷分布，由于LDA中包含了词语、主题和文本这三层结构，因此人们也将它叫做三层贝叶斯概率模型。在LDA模型中，计算机通过无监督的机器学习算法，从给定文本中推测它的主题分布。其中，每个文本是主题的概率分布，而每个主题又是词语的概率分布。

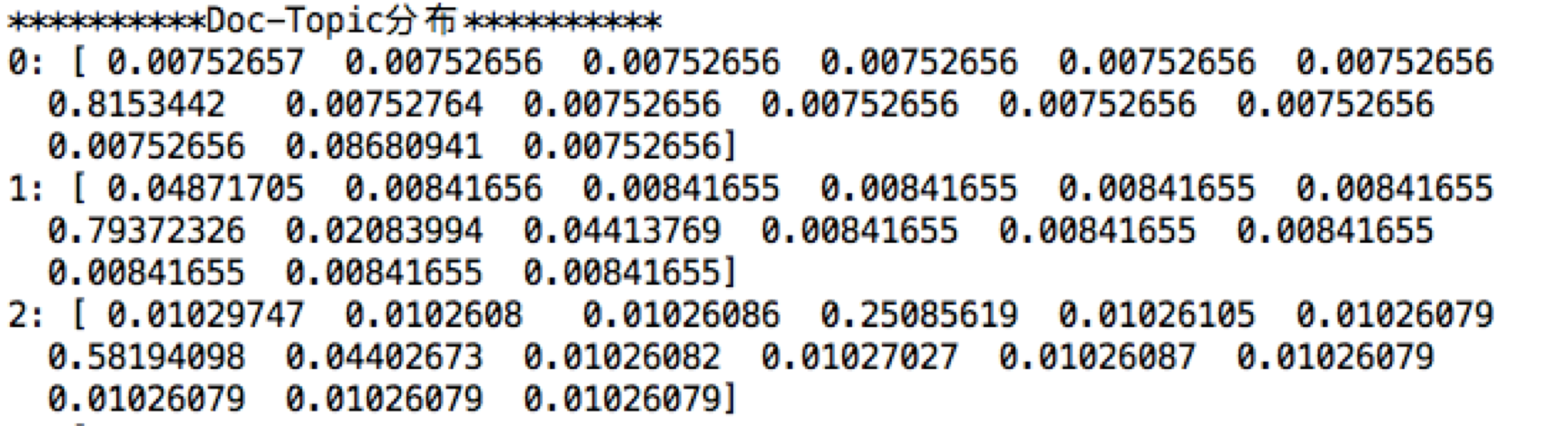


图 5.3 Doc-Topic分布

在上一节中我们利用K-means聚类分析了一段时间内数据中隐含的话题信息。本节我将利用LDA主题模型来对文本进行分析。同聚类一样，在LDA模型中，我们首先需要设置话题数，通过对文本进行建模，可以得到每篇文本所对应的主题分布。假设有n个主题，那么一篇文档的主题分布可以表示为，其中表示文档属于第k个话题的概率。也就是说，在LDA模型中，每一篇文档都可能对应多个主题，只是不同的主题概率不同。虽然人们在进行话题挖掘的时候经常使用这种方法，但是在本文中，由于待分析的数据都是食品安全话题下的，在对其进行建模的时候，分析结果并不具备很好的区分度。

## 5.3 实验与分析

表**5.1 K-means**提取话题结果

|  |  |  |
| --- | --- | --- |
|  | 2018年1-6月 | 2018年6-12月 |
| 0 | 相关 超市 执法人员 价格 餐饮 | 大闸蟹 螃蟹 养殖 水产 防伪 |
| 1 | 创建 示范 食品药品 餐饮 质量 | 食物 营养 健康 添加 食用 |
| 2 | 农产品 农业 出口 检验 种植 | 网络 餐饮 平台 外卖 订餐 |
| 3 | 食品药品 检查 重点 药品 违法 | 疫情 猪瘟 生猪 动物 防控 |
| 4 | 外卖 平台 餐饮 网络 订餐 | 批次 转载 合格 检出 超标 |
| 5 | 保健食品 宣传 虚假 整治 欺诈 | 检查 学校 重点 是否 餐饮 |
| 6 | 检验 检疫 进口 检疫局 入境 | 合格 抽检 批次 检验 监督 |
| 7 | 食物 样本 进口 残余 禽流感 | 农产品 农业 粮食 产业 蔬菜 |
| 8 | 食盐 盐业 定点 批发 价格 | 猪肉 月饼 检查 检疫 猪瘟 |
| 9 | 假冒 白酒 假酒 茅台 窝点 | 保健食品 宣传 虚假 欺诈 整治 |
| 10 | 配方 婴幼儿 奶粉 乳粉 乳业 | 检查 小作坊 加工 违法 添加剂 |
| 11 | 小作坊 酒类 白酒 加工 整治 | 相关 执法人员 现场 检查 超市 |
| 12 | 合格 批次 抽检 食用 超标 | 批次 合格 抽检 样品 食品药品 |
| 13 | 检查 餐饮 食堂 学校 学校食堂 | 食品药品 餐饮 重点 管理 全市 |
| 14 | 检测 农产品 农药 抽检 蔬菜 | 外卖 平台 商家 餐饮 料理 |

首先，我利用K-means聚类方法对2018年上半年和2018下半年关于食品安全数据进行了话题分析，在这里我将聚类数目设为15。通过表5.1，我们可以看出上半年比较突出的食品安全问题有食盐问题、假冒伪劣酒类问题、婴幼儿奶粉问题，下半年比较突出的问题有大闸蟹问题、猪肉疫病问题、月饼问题。而全年都存在的问题有食品药品、农产品质量问题、超市、餐饮外卖以及学校食堂的食品安全问题。通过对2018年食品安全数据的聚类分析，可以直观表现出热点话题，帮助人们了解当前最主要的食品安全问题。根据不同时间段的数据对比，人们也可以对食品安全问题进行预防和针对性的治理。

表**5.2 LDA**提取话题结果

|  |  |  |
| --- | --- | --- |
|  | 2018年1-6月 | 2018年6-12月 |
| 0 | 地沟油,核桃,假药,垃圾,师生 | 馒头,白糖,张店区,西瓜,船舶 |
| 1 | 曲奇,东阿,食责险,淡水鱼,皇冠 | 专案组,一桌,老窖,犯罪团伙,王某 |
| 2 | 龙口,该起,垃圾,齐河县,野生动物 | 池州,静安区,宝坻区,龙羊峡,大连 |
| 3 | 兔肉,四季豆,休耕,果酒,护苗 | 收运,减肥药,淄博市,新化县,绝味 |
| 4 | 发言人,本港,卫生署,禽肉,卫生组织 | 盒马,刘先生,老鼠,营养餐,罂粟 |
| 5 | 椰汁,海藻,豆浆,利津县,证明书 | 馒头,杨梅,泡菜,丙烯酰胺,阿根廷 |
| 6 | 食品,食品安全,餐饮,合格,检查 | 药酒,限塑令,泡制,榴莲,购物袋 |
| 7 | 旅客,谣言,转基因,燕窝,小龙虾 | 亚型,营口市,菜豆,黄石市, 栗子 |
| 8 | 金平,豆芽,核苷酸,主播,蛋奶 | 食品,合格,批次,食品安全,抽检 |
| 9 | 调和油,野菜,奶业,印尼,二审 | 龙井茶,馒头,槟榔,鱼胆,巧克力 |
| 10 | 潲水,底料,奶业,运城,乳企 | 三文鱼,虹鳟,寄生虫,一元,生食 |
| 11 | 除害,残余,规例,法例,鹤壁市 | 阳澄湖,大闸蟹,注水,咖啡因,葛粉 |
| 12 | 鲜奶,东莞,沙鱼,保健酒,黄酒 | 黑木耳,海底,乳饮料,香椿,干冰 |
| 13 | 淮北市,某合,茶垢,牛初乳,宁波市 | 疫情,封锁,生猪,非洲,猪瘟 |
| 14 | 大队长,潲水,金平,东阿,丙烯酰胺 | 样本,发言人,卫生署,跟进,规例 |

同样地，我利用LDA模型分析了2018年上半年和下半年的数据，设置话题数目为15。分析表5.2中的结果我们发现，2018年上半年话题6在每篇文章中所占的话题比例最高，2018年下半年话题8在每篇文章中所占的话题比例最高，而这两个话题的关键词都是“食品、食品安全”。每篇文章中其他话题所占的比例很小，几乎可以忽略不计。这是因为我们的数据集都是与食品安全相关的，每篇文本属于“食品”、“食品安全”话题下的概率就会很高。相应地，其他话题的概率就会很低，这就造成了我们无法从结果中分析出有价值的信息。因此，用LDA来进行话题分析不适用于本文中的数据，我们很难通过实验结果直观地看出话题类别。

## 5.4 本章小结

本章我介绍了通过文本聚类来实现话题提取的过程，通过使用K-means方法对向量化后的文本数据进行聚类，我们可以更加直观的看出一段时间内食品安全话题下的热点。另外，本章也对利用LDA模型来提取话题进行了实验，由于食品安全话题下的数据分类太过细小，因此这种方法对本文的话题提取并不适用。

# 第六章 结语

本章我将对论文内容进行总结并对之后的工作进行展望。

## 6.1 总结

本文对网络平台上的食品安全数据的采集技术进行了研究。通过观察和分析我发现网络平台上关于食品安全的数据通常为非结构化数据，包含了文本、图片、视频、音频等多种格式的文件。那么我们如何才能在这众多形式各异的数据中发现我们关注的话题成为了新的问题。本文通过爬虫来获取网络上的各类形式数据，然后将这些不同形式数据中都转换为文本，采用文本分类技术来进行建模，从而达到主题判别的目的，可以有效过滤垃圾信息。最后，本文使用了文本聚类技术来对这些文本数据进行聚类分析，从而使文本中隐含的话题更加明确和清晰。

本文的创新点主要有：

1.打破了传统对于食品安全领域研究仅限于文本形式网页的限制，能够对文本、音频、视频等多种格式的数据进行处理。

2.针对爬取到的数据中包含部分与食品安全话题无关的垃圾数据，使用文本分类方法来通过建模对文本的话题进行判别，能够有效过滤与食品安全话题无关的数据。

## 6.2 展望

虽然本文对食品安全数据进行了探讨和研究，对食品安全数据的采集实验也初步有了成果。但仍存在部分问题：

首先，在爬虫数据源的选择上，本文仅爬取了一定数目的网站进行爬取，爬取到的内容相对于整个网络中的数据来说只占很小的一部分，因此不论是对其进行分类建模还是聚类分析，都存在因为数据量太少造成实验结果不准确的问题。另外，在对视频进行字幕提取截取图像时，全部机械地截取图像下方部分区域，可能会造成一些无关文字也被截取进来，对文字识别的结果造成影响，从而影响视频文字内容反映的语义。在识别音频时，由于百度语音识别接口限制了音频长度为1分钟，在截取音频时，很可能会将连续的语音分割，造成后面识别结果不准确。其次，在对文本进行分类和聚类分析时，在参数的选择上没有进行细致的研究，会对实验结果造成影响。

因此在后续工作中，可以针对上述的三个问题进行改进。在爬虫部分，我需要改进爬虫的策略，爬取更多的数据源。针对视频字幕提取部分，为了替代现在这种机械确定字幕区域的方式，我需要对视频字幕的定位进行更深层次的研究。在进行音频分割时，可以考虑通过音频中句子之间的停顿来进行音频的截取，避免影响音频识别的效果。后期可以考虑改进文本处理方法，使用多种文本分类算法来进行主题判别，对比各种方法的效果。在利用文本聚类提取话题内容时，不仅需要对参数的选取做更多研究，还可以考虑选择除K-means外更好的聚类算法从而达到优化结果的目的。

# 参考文献

1. Gandomi A , Haider M . Beyond the hype: Big data concepts, methods, and analytics[J]. International Journal of Information Management, 2015, 35(2):137-144.
2. Gharehchopogh F S A , Khalifelu Z A B . Analysis and evaluation of unstructured data: Text mining versus natural language processing[C]. International Conference on Application of Information & Communication Technologies. IEEE, 2011.
3. 杨桥. 基于MongoDB的非结构化数据管理的研究与应用[D].电子科技大学,2017.
4. 王翠. 基于大数据的食品安全事件的研究[D].天津科技大学,2016.
5. 吴强强. 基于主题爬虫的食品安全网络舆情分析方法研究与监测系统开发[D].北京化工大学,2016.
6. 戴龙龙. 基于Storm分布式食品安全网络舆情监测方法研究[D].北京化工大学,2017.
7. 杨柳. 面向食药安全主题的突发话题检测技术研究[D].电子科技大学,2018.
8. 刘金硕, 彭映月, 章岚昕, et al. 网络食品安全问题话题发现的LDA-K-means算法[J]. 武汉大学学报(工学版), 2017(02):149-152.
9. 汪历伟. 食品供应链质量安全追溯服务平台研究与实现[D].电子科技大学,2016.
10. 林延昌. 基于区块链的食品安全追溯技术研究与实现[D].广西大学,2017.
11. Wang J , Yue H . Food safety pre-warning system based on data mining for a sustainable food supply chain[J]. Food Control, 2016:S0956713516305242.
12. Kate K , Chaudhari S , Prapanca A , et al. FoodSIS: a text mining system to improve the state of food safety in singapore[J]. 2014.
13. Marvin H J P , Esmée M. Janssen, Bouzembrak Y , et al. Big data in food safety: An overview[J]. Critical Reviews in Food Science and Nutrition, 2016, 57(11):2286-2295.
14. Ahearn M C , Armbruster W , Young R . Big Data's Potential to Improve Food Supply Chain Environmental Sustainability and Food Safety[J]. International Food and Agribusiness Management Review, 2016, 19.
15. Meyer C , Hamer M , Terlau W , et al. Web Data Mining and Social Media Analysis for better Communication in Food Safety Crises[J]. International Journal on Food System Dynamics, 2015.
16. Singh A , Shukla N , Mishra N . Social media data analytics to improve supply chain management in food industries[J]. Transportation Research Part E: Logistics and Transportation Review, 2017:S1366554516303817.
17. Abukausar M , S. Dhaka V , Kumar Singh S . Web Crawler: A Review[J]. International Journal of Computer Applications, 2013, 63(2):31-36.
18. Yu, Liu. Survey on topic-focused crawlers[J]. Computer Engineering & Science, 2015, 37(2): 231-237.
19. 陈晨. 面向Web文本挖掘的主题网络爬虫研究[D].电子科技大学,2017.
20. Zhang H , Zhao K , Song Y Z , et al. Text extraction from natural scene image: A survey[J]. Neurocomputing, 2013, 122(51):310-323.
21. Yang H, Meinel C. Content Based Lecture Video Retrieval Using Speech and Video Text Information[J]. IEEE Transactions on Learning Technologies, 2014, 7(2):142-154.
22. Bai J, Chen Z , Feng B , et al. Chinese Image Text Recognition on grayscale pixels[C]. IEEE International Conference on Acoustics. IEEE, 2014.
23. 黄翼彪. 开源中文分词器的比较研究[D].郑州大学,2013.
24. K. Nalini L. Jaba Sheela "Survey on Text Classification" International Journal of Innovative Research in Advanced Engineering vol. 1 no. 6 pp. 412-417 July 2014.

# 致谢

在完成毕业设计和论文的过程中，很多人给予了我帮助，在这里我要对他们表示衷心的感谢。

首先，我要感谢我的指导老师朱卫平老师，当我面对毕业设计一筹莫展的时候，是朱老师悉心教导，提供了一系列思路供我参考。在做毕业设计的过程中，朱老师定期让我们进行汇报，时刻关注我们的项目进展，并经常提出一些对我帮助很大的建议。这使得我能按时完成毕业设计，并在这个过程中学会了很多方法和道理。感谢老师无私的奉献和付出。其次，我还要感谢同学和朋友们的帮助，他们在我完成毕业设计的过程中一直不断地支持我，让我能够顺利完成毕业设计。最后我要感谢武汉大学，是学校浓厚的学术氛围感染影响了我，感谢学校的培养！