第11章 文本情感分析

第11章 文本情感分析

- 11.1 中文分词方法
- 11.2 文本的关键词提取
- 11.3 文本情感分析
- 11.4 LDA主题模型
- 11.5 运用LDA模型对电商手机评论进行主题分析

第11章 文本情感分析

互联网的快速发展，尤其Web2.0的出现，促进了互联网由“阅读式互联网”向“交互式互联网”转变。网络不仅成为人们获取信息的重要来源，也成为人们发表自己的观点和分享自己的体验，直接表达喜、怒、哀、乐等各种情感的重要平台，因而在网络中就形成了大量带有主观情感倾向性的文本。

通过电影或电视剧的评论分析，可以了解用户对节目的喜怒哀乐，进而制定好的剧情和上线时间；通过大众舆论导向分析，政府部门可以了解公民对热门事件的情感倾向，掌握大众舆论导向，为政府制定相关政策提供支持。

第11章 文本情感分析

- ✓ 11.1 中文分词方法
- ☐ 11.2 文本的关键词提取
- ☐ 11.3 文本情感分析
- ☐ 11.4 LDA主题模型
- ☐ 11.5 运用LDA模型对电商手机评论进行主题分析

11.1 中文分词方法

- 把中文的汉字序列切分成有意义的词，就是中文分词，也称为切词。
- 中文分词具有以下几个方面的词认定问题，
  - 一是词的界定模糊，中文处于一个不断变化的过程，词也一样，一直有新词源源不断地创作出来，但新词是否被认同有一个过程，这就造成有些词只被部分人认定成词；
  - 二是长词问题，如“中国人民解放军”从功能上看是一个整体，我们可以将其当做一个词，但也可以切分成“中国人民 / 解放军”，将其看做组合词；
  - 三是词的扩展问题，有些词由于自身的特性，可以添加一些量词来进行扩展，如“吃饭”可以说成“吃了一点饭”、“上课”说成“上了节课”，词的形式的多样性使得词的界定比较困难。

11.1 中文分词方法

11.1.1 基于字符串匹配的分词方法

（1）正向最大匹配

- 正向最大匹配算法假设分词词典中最长词条所含的汉字个数是MaxLen，每次从待切分字串S的开始处截取一个长度为MaxLen的字串W，将W同词典中长度为MaxLen的词条依次相匹配，如果某个词条与其完全匹配，则把W作为一个词从S中切分出去，然后再从S的开始处截取另一个长度为MaxLen的字串，重复与词典中词条相匹配的过程，直到待切分字符串为空。
- 如果在词典中找不到长度为MaxLen词条与W匹配，就从W的尾部减去一个字，用剩下的长度为MaxLen-1的字符串与词典中长度为MaxLen-1词条进行匹配，如果匹配成功则将该长度为MaxLen-1的字串切分出去，否则再从W尾部减去一个字，重复匹配过程，直到匹配成功。

11.1 中文分词方法

11.1.1 基于字符串匹配的分词方法

（1）正向最大匹配

假设要进行分词的字串为“研究生命的起源”。假定字典包含的词条如下：研究、研究生、生命、命、的、起源、

假定最大匹配字数设定为5，正向最大匹配过程：

第1轮扫描：

第1次：“研究生命的”，扫描5字词典，无匹配。

第2次：“研究生命”，扫描4字词典，无匹配。

第3次：“研究生”，扫描3字词典，匹配成功。

扫描中止，输出的第1个词为“研究生”，去除第1个词后开始第2轮扫描，即：

11.1 中文分词方法

11.1.1 基于字符串匹配的分词方法

（1）正向最大匹配

假设要进行分词的字串为“研究生命的起源”。

第2轮扫描：

第1次：“命的起源”，扫描4字词典，无匹配。

第2次：“命的起”，扫描3字词典，无匹配。

第3次：“命的”，扫描2字词典，无匹配。

第4次：“命”，扫描1字词典，匹配成功。

扫描中止，输出的第2个词为“命”，去除第2个词后开始第3轮扫描，即：

11.1 中文分词方法

11.1.1 基于字符串匹配的分词方法

（1）正向最大匹配

假设要进行分词的字串为“研究生命的起源”。

第3轮扫描：

第1次：“的起源”，扫描3字词典，无匹配。

第2次：“的起”，扫描2字词典，无匹配。

第3次：“的”，扫描1字词典，匹配成功。

扫描中止，输出的第3个词为“的”，去除第3个词后开始第4轮扫描，即：

第1次：“起源”，扫描2字词典，匹配成功。

正向最大匹配法，最终切分结果为：研究生/命/的/起源。

11.1 中文分词方法

11.1.2 基于统计的分词方法

从形式上看，词是稳定的字的组合，因此在上下文中，如果相连的字在不同的文本中出现的次数越多，就证明这些相连的字很可能就是一个词。因此，可以利用字与字相邻共现的频率或概率来反映成词的可信度。可以对语料中相邻共现的各个字的组合的频度进行统计，当组合频度高于某一个临界值时，便可认为此字组可能会构成一个词。

第11章 文本情感分析

- 11.1 中文分词方法
- ✓ 11.2 文本的关键词提取
- 11.3 文本情感分析
- 11.4 LDA主题模型
- 11.5 运用LDA模型对电商手机评论进行主题分析

11.2 文本的关键词提取

在自然语言处理领域，一个关键的问题是关键词（也称为特征）的提取，关键词提取的好坏将会直接影响算法的效果，常用的关键词提取方法有文档频率、互信息、词频-逆文件频率（TF-IDF）等。

11.2.1 基于文档频率的关键词提取

在文档频率方法中，将包含关键词的文档的数目和所有文档数目的比值作为文档频率。在进行关键词抽取的时候，首先计算出每一个关键词的文档频率，然后设定合适的阈值，根据阈值去进行关键词选取。如果文档频率低于某一阈值，说明该关键词在文档中占的权重相对较弱，就丢弃，但如果文档频率大于某一值，说明该关键词在文档中出现的比较频繁，不具有代表性，也会被丢弃，最后剩下的就是需要的关键词。

11.2 文本的关键词提取

11.2.2 基于互信息的关键词提取

互信息是信息论中描述事件A和事件B同时出现，发生相关联而提供的信息量。在分类问题中，可以使用互信息衡量某一个特征和特定类别的相关性，互信息越大，说明该特征和该特定类别的相关性越大，反之相关性越小。因而，互信息可以有效地体现特征与文本类别的关联度。

特征词t与文本类别C之间的互信息定义为：

$$MI(t,C)=\log_2\frac{P(t,C)}{P(t)P(C)}$$

其中，P(t, C)表示类别C中包含特征词t的文本数与总文本数的比率，P(t)表示出现特征词t的文本数与总文本数的比率，P(C)表示属于类别C的文本数与总文本数的比率。

11.2 文本的关键词提取

11.2.3 基于词频-逆文件频率的关键词提取

词频-逆文件频率TF-IDF（term frequency–inverse document frequency）的基本思想是：词语的重要性与它在文件中出现的次数成正比，但同时会随着它在语料库中出现的频率成反比下降。也就是说，一个词语w在一篇文章d中出现次数越多，并且在其他文档中很少出现，则认为词语w具有很好的区分能力，该词语与文章d的相关程度就越高，越能够代表该文章，适合用来把文章d和其他文章区分开来。

- 词频（term frequency，TF）指的是一个词在文件中出现的频率。
- 逆向文件频率（inverse document frequency，IDF）用来衡量某一词语在文件集的重要性，某一特定词语的IDF，可以由文件集的总文件数目除以包含该词语的总文件数目，再将得到的商取对数得到。

11.2 文本的关键词提取

11.2.3 基于词频-逆文件频率的关键词提取

对文件集D，设|D|表示D中的总文件数，$|D_i|$表示D中含有第i种词的总文件数，用IDF_i表示第i种词在文件集D的逆行文件频率，则IDF_i定义为：

$$IDF_i = \log \frac{|D|}{|D_i|+1}$$

利用文件内的较高的词语频率，以及词语在整个文件集合中的较低文件频率，可以得到较高权重的TF-IDF词语，这些词语在该文件中具有较高的重要程度。因此，以TF和IDF的乘积作为选取特征词的测度，通过TF-IDF选取重要词语可用于过滤掉常见的词语，得到重要的词语。

11.2 文本的关键词提取

11.2.3 基于词频-逆文件频率的关键词提取

下面给出利用jieba（结巴分词）系统中的TF-IDF实现中文文本的关键词抽取。

```
>>> from jieba import analyse
>>> tfidf = analyse.extract_tags
>>> text = "进程是计算机中的程序关于某数据集合上的一次运行活动，是系统进行资源分配和调度的基本单位，是操作系统结构的基础。在早期面向进程设计的计算机结构中，进程是程序的基本执行实体；在当代面向线程设计的计算机结构中，进程是线程的容器。程序是指令、数据及其组织形式的描述，进程是程序的实体。"
```

11.2 文本的关键词提取

11.2.3 基于词频-逆文件频率的关键词提取

```
#基于TF-IDF算法从text文本抽取10个关键词并返回关键词权重
>>> keywords = tfidf(text,topK = 10,withWeight = True)
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\caojie\AppData\Local\Temp\jieba.cache
Loading model cost 0.995 seconds.
Prefix dict has been built succesfully.
>>> print(keywords)
[('进程', 0.679731445205208 3), ('程序', 0.527734587209166 7), ('线程', 0.498115312620833 3), ('计算机', 0.42529902693), ('面向', 0.313145135540833 34), ('结构', 0.3019362256125), ('实体', 0.281544168719583 3), ('资源分配', 0.23057217308125), ('设计', 0.230296829615), ('基本', 0.2000482131291666 5)]
```

第11章 文本情感分析

- 11.1 中文分词方法
- 11.2 文本的关键词提取
- ✓ 11.3 文本情感分析
- 11.4 LDA主题模型
- 11.5 运用LDA模型对电商手机评论进行主题分析

11.3 文本情感分析

文本情感分析是指用自然语言处理、文本挖掘以及计算机语言学等方法来识别和提取原素材中的主观信息。目的是为了找出文本中作者对某个实体（包括产品、服务、人、组织机构、事件、话题）的评判态度（支持或反对、喜欢或厌恶等）或情感状态（高兴、愤怒、悲伤、恐惧等）。

11.3.1 文本情感分析的层次

文本情感倾向分析可以分成词语情感倾向性分析、句子情感倾向性分析、文档情感倾向性分析、海量信息的整体倾向性预测四个层次。

1. 词语情感倾向性分析

词语情感倾向分析包括对词语极性、强度和上下文模式的分析。词语情感倾向分析目前主要有三种方法：

11.3 文本情感分析

11.3.1 文本情感分析的层次

1. 词语情感倾向性分析

词语情感倾向分析目前主要有三种方法：

（1）由已有的词典或词语知识库生成情感倾向词典。

该方法通过给定一组已知极性的词语集合作为种子，对于一个情感倾向未知的新词，在词典中找到与该词语义相近、并且在种子集合中出现的若干个词，根据这几个种子词的极性，对未知词的情感倾向进行推断。

（2）无监督机器学习的方法。该方法假设已经有一些已知极性的词语作为种子词，对于一个新词，根据词语在语料库中的同现情况判断其联系紧密程度。

11.3 文本情感分析

11.3.1 文本情感分析的层次

1. 词语情感倾向性分析

（3）基于人工标注语料库的学习方法。首先对情感倾向分析语料库进行手工标注，标注的级别有文档级的情感倾向性、短语级的情感倾向性和分句级的情感倾向性。在这些语料的基础上，在大规模语料中利用词语的共现关系、搭配关系或者语义关系，判断其它词语的情感倾向性。

11.3 文本情感分析

11.3.1 文本情感分析的层次

2. 句子情感倾向性分析

词语情感倾向分析的处理对象是单独的词语，而句子情感倾向性分析的处理对象则是在特定上下文中出现的语句。其任务就是对句子中的各种主观性信息进行分析和提取，包括对句子情感倾向的判断，以及从中提取出与情感倾向性论述相关联的各个要素，包括情感倾向性论述的持有者、评价对象、倾向极性、强度，甚至是论述本身的重要性等。

11.3 文本情感分析

11.3.1 文本情感分析的层次

3. 文档情感倾向性分析

文档级情感分析旨在从整体上判断某个文本的情感倾向性。代表性的工作是Turney和Pang对电影评论的分类。Turney的方法是将文档中词的倾向性进行平均，来判断文档的倾向性。这种方法基于情感倾向性词典，不需要人工标注文本情感倾向性的训练语料。Pang的任务是对电影评论的数据按照倾向性分成两类，他利用人工标注了文本倾向性的训练语料，基于一元分词（把句子分成一个一个的汉字）和二元分词（把句子从头到尾每两个字组成一个词语）等特征训练分类器，通过训练的分类器实现对电影评论的倾向性分类。

4. 海量信息的整体倾向性预测

海量信息的整体倾向性预测的主要任务是对从不同信息源抽取出的、针对某个话题的情感倾向性信息进行集成和分析，进而挖掘出态度的倾向性和走势。

11.3 文本情感分析

11.3.2 中文文本情感倾向分析

情感分析就是分析一句话说得是主观还是客观描述，分析这句话表达的是积极的情绪还是消极的情绪。下面通过“这手机的画面极好，操作也比较流畅。不过拍照真的太烂了！系统也不好。”来阐述中文文本情感倾向分析。

（1）分析句子中的情感词

要分析一句话是积极的还是消极的，最简单最基础的方法就是找出句子里面的情感词。出现一个积极的情感词，情感分值就加1，出现一个消极的情感词，情感分值就减1。

句子中就有“好”，“流畅”两个积极情感词，“烂”一个消极情感词，其中“好”出现了两次，句子的情感分值就是1+1-1+1=2。

11.3 文本情感分析

11.3.2 中文文本情感倾向分析

下面通过“这手机的画面极好，操作也比较流畅。不过拍照真的太烂了！系统也不好。”来阐述中文文本情感倾向分析。

（2）分析句子中的程度词

“好”，“流畅”和“烂”前面都有一个程度修饰词。“极好”就比“较好”或者“好”的情感更强，“太烂”也比“有点烂”情感强得多。所以需要在找到情感词后往前找一下有没有程度修饰词，并给不同的程度修饰词一个权值。比如有“极”，“无比”，“太”程度词，就把情感分值乘4；有“较”，“还算”程度词，就把情感分值乘2，有“只算”、“仅仅”这些程度词，就乘0.5了。考虑到程度词，句子的情感分值就是：1*4+1*2-1*4+1=3。

（3）分析句子中的感叹号

可以发现太烂了后面有感叹号，叹号意味着情感强烈。因此发现叹号可以为情感值加2（正面的）或减2（负面的）。考虑到感叹号， 句子的情感分值就变成了：4*1+1*2-1*4-2+1=-1

11.3 文本情感分析

11.3.2 中文文本情感倾向分析

下面通过“这手机的画面极好，操作也比较流畅。不过拍照真的太烂了！系统也不好。”来阐述中文文本情感倾向分析。

（4）分析句子中的否定词

最后面那个“好”并不是表示“好”，因为前面还有一个“不”字。所以在找到情感词的时候，需要往前找否定词。比如“不”，“不能”、“非”“否”这些词。而且还要数这些否定词出现的次数，如果是单数，情感分值就乘(-1)，但如果是偶数，那情感就没有反转，保持原来的情感分值。在这句话里面，可以看出“好”前面只有一个“不”，所以“好”的情感值应该反转，乘(-1)。这时候，这句话的准确情感分值变为

4*1+1*2-1*4-2+1*(-1)= -1

11.3 文本情感分析

11.3.2 中文文本情感倾向分析

下面通过“这手机的画面极好，操作也比较流畅。不过拍照真的太烂了！系统也不好。”来阐述中文文本情感倾向分析。

（5）积极和消极分开来分析

很明显就可以看出，这句话里面有褒有贬，不能仅用一个分值来表示它的情感倾向。此外，权值的设置方式也会影响最终的情感分值。因此，对这句话恰当的处理是给出一个积极分值、一个消极分值，这样消极分值也是正数，无需使用负数了。它们同时代表了这句话的情感倾向，这时候句子的情感分值就表示为：积极分值为6，消极分值为7。

（6）以分句的情感为基础进行情感分析

再细分一下，一条评论的情感分值是由不同的分句决定的，因此要得到一条评论的情感分值，就需要先计算出评论中每个句子的情感分值。前面列举的评论有四个分句，以分句的情感为基础进行情感分析，评论的情感分值结构变为[[4, 0], [2, 0], [0, 6], [0, 1]]，列表中的每个子列表中的两个值一个表示分句的积极分值一个表示分句的消极分值。

第11章 文本情感分析

- 11.1 中文分词方法
- 11.2 文本的关键词提取
- 11.3 文本情感分析
- ✓ 11.4 LDA主题模型
- 11.5 运用LDA模型对电商手机评论进行主题分析

https://www.bilibili.com/video/BV1LQ4y1Q7xv?p=1

https://www.bilibili.com/video/BV1Sr4y1C7Xc/?spm_id_from=333.788

11.4 LDA主题模型

- 潜在狄立克雷分配模型LDA（Latent Dirichlet Allocation）是概率生成性模型的一个典型代表，也将其称为LDA主题模型。
- 所谓生成模型，就是说一篇文章的每个词可通过“以一定概率选择某个主题，并从这个主题中以一定概率选择某个词语”这样的过程而得到。
- 所谓“主题”就是一个文本所蕴含的中心思想，一个文本可以有一个主题，也可以有多个主题。
- 主题由关键词来体现，可以将主题看作是一种关键词集合，同一个词在不同的主题背景下，它出现的概率是不同的，如一篇文章出现了某个球星的名字，我们只能说这篇文章有很大概率属于体育的主题，但也有小概率属于娱乐的主题。
- 可以说，LDA用词汇的分布来表达主题，将主题看作是一种词汇分布，用主题的分布来表达文章。LDA把文章看作是由词汇组合而成，LDA通过不同的词汇概率分布来反映不同的主题。一组词汇越能反映主题，这组词汇整体的出现概率越大。

11.4 LDA主题模型

- 举例说明LDA通过词汇的概率分布来反映主题。

假设有词汇集合{乔丹，篮球，足球，奥巴马，克林顿}，假设有两个主题{体育，政治}。LDA认为体育这个主题具有：{乔丹:0.3，篮球:0.3，足球:0.3，奥巴马:0.02克林顿:0.03}，其中数字代表词出现的概率。而政治这个主题有：{科比:0.03，篮球:0.03，足球:0.04，奥巴马:0.3，克林顿:0.3}。

- 举例说明LDA通过主题的分布来表达文章。

假设现在有两篇文章《体育快讯》和《娱乐周报》，有三个主题“体育”，“娱乐”，“废话”。LDA认为《体育快讯》是这样的{废话:0.1, 体育:0.7, 娱乐:0.2}，而《娱乐周报》是这样的{废话:0.2，娱乐:0.7, 体育:0.1}。也就是说，一篇文章在讲什么，通过不同的主题比例就可以得出。

11.4 LDA主题模型

- 总的来说，LDA认为每个主题对应一个词汇分布，而每个文档会对应一个主题分布。一篇文章的生产过程是这样的：

（1）确定主题和词汇的分布；

（2）确定文章和主题的分布；

（3）随机确定该文章的词汇个数N；

（4）如果当前生成的词汇个数小于N执行第5步，否则执行第6步；

（5）由文档和主题分布随机生成一个主题，通过该主题由主题和词汇分布随机生成一个词，继续执行第4步；

（6）文章生成结束。

- 只要确定好两个分布（主题与词汇分布，文章与主题分布），然后随机生成文章各个主题比例，再根据各个主题随机生成词，忽略词与词之间的顺序关系，这就是LDA生成一篇文章的过程。

第11章 文本情感分析

- ☐ 11.1 中文分词方法
- ☐ 11.2 文本的关键词提取
- ☐ 11.3 文本情感分析
- ☐ 11.4 LDA主题模型
- ✓ 11.5 运用LDA模型对电商手机评论进行主题分析

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.1 电商手机评论数据的采集

- 要分析电商平台的手机评论数据，首先需要对评论数据进行采集，这里采用网络爬虫工具进行数据采集。网上比较流行的免费采集器有这么几个：火车头，海纳，ET，三人行，八爪鱼，狂人。下面，采用八爪鱼采集器采集手机评论数据。
- 八爪鱼采集的核心原理是：模拟人浏览网页，复制数据的行为，通过记录和模拟人的一系列上网行为，代替人眼浏览网页，代替人手工复制网页数据，从而实现自动化从网页采集数据，然后通过不断重复一系列设定的动作流程，实现全自动采集大量数据。

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.2 评论预处理

查看采集到的iPhoneX评论文本后，可以发现评论具有以下特点。

（1）文本短，很多评论就是一句话。

（2）情感倾向明显，如“好”、“可以”、“漂亮”。

（3）语言不规范，会出现一些网络用词、符号、数字等，如“666”、“神器”；

（4）重复性大，一句话出现多次词语重复，如“很好，很好，很好”。

- 总的来说，文本评论数据里存在大量价值含量很低甚至没有价值的评论，如果对这些评论数据进行分词、词频统计、提取主题乃至情感分析，必然造成很大的干扰，评论数据分析结果的质量也将会受到很大的影响。因此，在利用这些评论文本进行数据分析之前就必须对文本进行预处理，去除低价值、无价值的评论。
- 文本评论数据的预处理主要包括3个方面：文本去重、机械压缩去词、短句删除。

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.2 评论预处理

对评论文本去重的算法实现

```
>>> import pandas as pd
>>> import re
#读取采集的iPhoneX评论数据
>>>Data=pd.read_csv("C:/Users/iPhoneX_comment.csv",sep =",",encoding = "utf-8")
>>>Data["页面标题"].value_counts()#返回包含值和该值出现的次数
Apple 苹果iPhoneX 手机 银色 64G 标配【图片 价格 品牌 报价】-京东    1000
Name: 页面标题, dtype: int64
#删除冗余属性，返回一个新对象
>>> iPhoneX_comment_new = Data.drop(labels=["会员","级别","评价星级","时间","点赞数","评论数","追评时间","追评内容","页面网址","页面标题","采集时间"],axis=1)
```

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.2 评论预处理

对评论文本去重的算法实现

```
>>> iPhoneX_comment_new.head(2)  #获取删除冗余属性后的前两条记录
                                        评价内容
0  非常快 很好的体验 还好贴心的送来贴膜 手机透明壳壳五星好评
1  运了三天才到～今早着急的我亲自跑去站点取 非常好  颜色也很美 沉甸甸的很庆幸买了这家 虽...
>>> iPhoneX_comment_new.duplicated().sum()  #统计重复的文本数
6
#去除6行重复评论
>>> iPhoneX_comment_unique = iPhoneX_comment_new.drop_duplicates()
>>> iPhoneX_comment_unique.shape
(994, 1)
#编译正则表达式对象，用于去除高频无意义的词
>>> strinfo = re.compile("手机|苹果|店家|京东|东西|n")
>>>    iPhoneX_comment_useless    =    iPhoneX_comment_unique["评价内容
"].apply(lambda x:strinfo.sub("",x))
```

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.3 评论文本分词

- 中文分词指的是将一个汉字序列切分成一个一个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。分词是分析文本评论的关键步骤，只有分词准确，才能得到正确的词频，也才能通过词频-逆文件频率（TF-IDF）提取到正确的关键词。如果分词效果不佳，即使后续算法优秀也无法实现理想的效果。例如在特征选择的过程中，不同的分词效果，将直接影响词语在文本中的重要性，从而影响特征的选择。
- 下面采用利用jieba分词包对评论文本进行中文分词。

```
import jieba
wordsCut=Conment["iPhoneX评论"].astype("str").apply(lambda x: list(jieba.cut(x)))
wordsCut[:2]
```

运行上述代码得到输出:

```
0    [非常, 快, , 很, 好, 的, 体验, , 还好, 贴心, 的, 送来, 贴膜, ...
1    [运了, 三天, 才, 到, ～, 今早, 着急, 的, 我, 亲自, 跑, 去, 站点, ...
```

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.4 去除停用词

停用词是一些完全没有用或者没有意义的词，停用词大致可分为如下两类：

1、使用十分广泛，甚至是过于频繁的一些单词。比如英文的“i”、“is”、“what”，中文的“我”、“就”之类词几乎在每个文档上均会出现

2、文本中出现频率很高，但实际意义又不大的词。这一类主要包括了语气助词、副词、介词、连词等，通常自身并无明确意义，只有将其放入一个完整的句子中才有一定作用的词语。如常见的“的”、“在”、“和” 之类。

```
#加载停用词表，sep设置为文档内不包含的内容，否则会出错
>>> stopWords = pd.read_csv("stoplist.txt",sep = "fenci",encoding ="utf-8",header = None)
>>> stopWords.head(2)
   0
0  说
1  人
>>> stopWords = list(stopWords[0]) + [" ", ""]  #向stopWords里添加空格符
```

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.4 去除停用词

```
#去除停用词
>>> wordStop = wordsCut.apply(lambda x:[i for i in x if i not in stopWords])
>>> wordStop[:10]     #显示去除停用词后的前10个文本
0                [体验, 还好, 贴心, 送来, 贴膜, 透明, 壳, 壳, 五星, 好评]
1    [运了, 三天, 今早, 着急, 跑, 站点, 取, 颜色, 美, 沉甸甸, 庆幸, 这家,...
2    [说实话, 拿到, 手, 惊艳, 6sp, 机身, 小巧, 点, 屏幕, 倒, 刘海, 屏幕...
3    [感觉, 慢慢, 琢磨, 这家, 店, 不好, 感觉, 包装, 安心, 物流, 慢, 不怪,...
4    [自营, 抢到, 发货, 这家, 担心, 上海, 仓, 发货, 西南, 远, 第三天, 中午...
5                        [到手, 值得, 信赖, 期待已久, 终于, 到货]
6                              [物流, 速度, 接受, 很快]
7          [物流, 特快, 服务, 周到, 耐心, 解答, 拿到, 心情, 质量, 没得说, 超值]
8                          [原装, 查, 保修, 日期, 配件, 真假]
9    [比官, 网, 便宜, 三百, 三十多, 昨晚, 六点, 下单, 今天上午, 十点, 快递,...
Name: iPhoneX评论, dtype: object
```

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.5 绘制评论文本的词云图

词云图又叫文字云，是对文本数据中出现频率较高的关键词予以视觉上的突出，形成“关键词的渲染”，使人一眼就可以领略文本数据的主要表达意思。从技术上来看，词云是一种数据可视化方法，互联网上有很多的现成的工具：

（1）Tagxedo可以在线制作个性化词云。

（2）Tagul是一个Web服务，同样可以创建华丽的词云。

（3）Tagcrowd 还可以输入web的url，直接生成某个网页的词云。

（4）wordcloud是Python的一个第三方模块，使用wordcloud下的WordCloud函数生成词云。

下面给出绘制评论文本的词云图的代码实现。

```
from imageio import imread
from wordcloud import WordCloud,ImageColorGenerator
import matplotlib.pyplot as plt
wordTemp = []
```

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.5 绘制评论文本的词云图

下面给出绘制评论文本的词云图的代码实现。

```
wordTemp = []
for i in wordStop.index:
    wordTemp.extend(wordStop.loc[i])
wordStop_df = pd.DataFrame(wordTemp)
wordStop_df.columns = ["words"]
result= "/".join(wordTemp)
plt.rcParams['figure.figsize'] = (10.0, 10.0)
#自定义词云背景图片
image =imread("D:\\Python\\tupian.jpg")
#构建词云模型
wordcloud = WordCloud(background_color="white",
mask=image,font_path=r"C:\Windows\Fonts\simhei.ttf", max_font_size=200)
wordcloud.generate(result)
image_color=ImageColorGenerator(image)#从背景图片生成词云图中文字的颜色
```

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.5 绘制评论文本的词云图

下面给出绘制评论文本的词云图的代码实现。

```
image_color=ImageColorGenerator(image)#从背景图片生成词云图中文字的颜色
wordcloud.recolor(color_func=image_color)
#保存绘制好的词云图，比直接程序显示的图片更清晰
wordcloud.to_file(r"D:\Python\w
#显示词云图图片
plt.figure("词云图")  #指定所绘
plt.imshow(wordcloud)  #以图片
plt.axis("off")       # 关闭图像坐
plt.show()
```

绘制的词云图如图所示。

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.6 评论文本情感倾向分析

- 评论文本情感倾向分析主要是基于用户评论信息来分析出用户对某个特定事物的观点、看法、情感倾向以及情感色彩。通过对商品评论挖掘，商家可以及时的获取用户的需求和关注点，了解产品的不足之处，以便及时调整销售策略，实现精准营销，节约企业成本。
- 这里采用最简单的情感分析方法，即基于情感词典的情感分析方法，使用情感词典进行情感分析的主要思路是：对文档分词，找出文档中的情感词、否定词以及程度副词，然后判断每个情感词之前是否有否定词及程度副词，将它之前的否定词和程度副词划分为一个组，如果有否定词将情感词的情感权值乘以-1，如果有程度副词就乘以程度副词的程度值，最后所有组的得分加起来，大于0的归于正向，小于0的归于负向。

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.6 评论文本情感倾向分析

使用snownlp模块可以方便地直接对评论文本进行情感分析，将评论文本分为正面评论文本和负面评论文本，代码实现如下。

```
import pandas as pd
from snownlp import SnowNLP
#读取采集的iPhoneX评论数据
Data = pd.read_csv("C:/Users/iPhoneX_comment.csv", usecols=[3], header=0)
Data.duplicated().sum() #统计重复的文本数
#去除重复行评论
iPhoneX_comment_unique = Data.drop_duplicates()
coms=[]
coms=iPhoneX_comment_unique['评价内容'].apply(lambda x: SnowNLP(x).sentiments)
```

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.6 评论文本情感倾向分析

```
#情感分析，coms在0 1之间，以0.5分界，大于0.5为正面情感
pos_data=iPhoneX_comment_unique[coms>=0.6]   #取0.6是为了使情感更强烈
neg_data=iPhoneX_comment_unique[coms<0.4]   #获取负面情感评论文本数据集
print('正面评论文本的前3条记录：\n',pos_data[:3])
print('负面评论文本的前3条记录：\n',neg_data[:3])
```

运行上述代码得到的输出结果如下：

正面评论文本的前3条记录：

	评价内容
0	非常快 很好的体验 还好贴心的送来贴膜 手机透明壳壳五星好评
1	运了三天才到～今早着急的我亲自跑去站点取 非常好 颜色也很美 沉甸甸的 很庆幸买了这家 虽...
2	说实话拿到手很惊艳，跟6sp比起来，机身更小巧了点，但是屏幕倒大的多，刘海已经不重要了，屏幕...

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.6 评论文本情感倾向分析

负面评论文本的前3条记录：

	评价内容
8	好就是好，没的说，是原装，查了保修日期，就是配件不知道真假
9	比官网便宜三百三十多，昨晚六点多下单，今天上午十点京东快递送到，非常满意，官网太慢要退货了。...
14	非常好 手机很完美 今年应该是最后一样剁手的东西的，完美收官，自营的我是plus会员也没抢到...

- 接下来对正面评论文本数据集pos_data和负面评论文本数据集neg_data进行jieba分词和去除停用词，这两个内容已经在评论文本分词和去除停用词章节讲述了，此处不再赘述，最后得到正面评论文本情感分词集和负面评论文本情感分词集。

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.7 评论文本的LDA主题分析

评论文本的正面评价和负面评价混淆在一起，直接进行LDA主题分析可能会在一个主题下生成一些令人迷惑的词语，因而，应分别对正面评价和负面评价两类文本进行LDA主题分析。

下面使用Python开源的第三方Gensim库完成LDA主题分析。Gensim支持包括TF-IDF，LSA，LDA，和word2vec在内的多种主题模型算法。Gensim的基本概念：

- 语料（Corpus）：一组原始文本的集合，这个集合是Gensim的输入，Gensim会从这个语料中推断出它的结构，主题等。在Gensim中，Corpus通常是一个可迭代的对象（比如列表）。
- 向量（Vector）：由一组文本特征构成的列表，是一段文本在Gensim中的内部表达。
- 稀疏向量（Sparse Vector）：通常，我们可以略去向量中多余的0元素。此时，向量中的每一个元素是一个(key, value)的元组。
- 模型（Model）：是一个抽象的术语。定义了两个向量空间的变换（即从文本的一种向量表达变换为另一种向量表达）。

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.7 评论文本的LDA主题分析

下面给出评论文本的LDA主题分析的代码实现。

```
from gensim import corpora, models
pos = feelScore.loc[feelScore['Score'] > 0,'iPhoneX评论']
neg = feelScore.loc[feelScore['Score'] < 0,'iPhoneX评论']
#负面主题分析
neg_dict = corpora.Dictionary(neg)              #建立负面词典
neg_corpus = [neg_dict.doc2bow(i) for i in neg] #建立负面语料库
```

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.7 评论文本的LDA主题分析

```
#构建LDA模型
neg_lda = models.LdaModel(neg_corpus, num_topics = 3, id2word = neg_dict)
print("\n负面评价")
for i in range(3):
    print("主题%d : " %i)
    print(neg_lda.print_topic(i) ) #输出主题
#正面主题分析
pos_dict = corpora.Dictionary(pos)
pos_corpus = [pos_dict.doc2bow(i) for i in pos]
pos_lda = models.LdaModel(pos_corpus, num_topics = 3, id2word = pos_dict)
print("\n正面评价")
for i in range(3):
    print("主题%d : " %i)
    print(pos_lda.print_topic(i) ) #输出主题
```

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.7 评论文本的LDA主题分析

运行上述代码得到的输出结果如下：

负面评价

主题0：

0.019*"发现" + 0.013*"贵" + 0.011*"客服" + 0.010*"重启" + 0.010*"暂时" + 0.010*"降价" + 0.010*"习惯" + 0.009*"太" + 0.009*"发货" + 0.009*"适配"

主题1：

0.017*"舒服" + 0.017*"毛病" + 0.016*"屏幕" + 0.014*"物流" + 0.011*"感觉" + 0.011*"不错" + 0.008*"x" + 0.008*"发货" + 0.008*"帮" + 0.008*"朋友"

主题2：

0.015*"慢" + 0.014*"发现" + 0.012*"找" + 0.012*"真的" + 0.011*"充电" + 0.011*"充电器" + 0.011*"换" + 0.011*"坏" + 0.011*"快递" + 0.009*"发货"

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.7 评论文本的LDA主题分析

运行上述代码得到的输出结果如下：

正面评价

主题0：

0.040*"不错" + 0.020*"正品" + 0.013*"发货" + 0.013*"收到" + 0.009*"满意" + 0.009*"价格" + 0.008*"喜欢" + 0.007*"感觉" + 0.007*"便宜" + 0.007*"包装"

主题1：

0.020*"发货" + 0.018*"不错" + 0.013*"快递" + 0.012*"物流" + 0.010*"很快" + 0.010*"速度" + 0.010*"慢" + 0.009*"价格" + 0.008*"正品" + 0.007*"喜欢"

主题2：

0.033*"正品" + 0.028*"不错" + 0.020*"发货" + 0.017*"速度" + 0.015*"收到" + 0.015*"很快" + 0.013*"物流" + 0.013*"国行" + 0.010*"满意" + 0.008*"快递"

11.5 运用LDA模型对电商手机评论进行主题分析

11.5.7 评论文本的LDA主题分析

经过LDA主题分析后，正面评价和负面评价评论文本分别被聚成3个主题，每个主题下显示10个最有可能出现的词语以及相应的概率。

- 根据对iPhoneX好评的3个潜在主题的特征词提取，主题0中的高频特征词有“不错”、“正品”、“发货”、“收到”、“满意”、“价格”、“喜欢”等，主要反映京东上的iPhoneX质量不错，是正品，值得购买；主题1中的高频特征词有“发货”、“不错”、“快递”、“物流”、“很快”、“速度”等，主要反映京东的发货、物流速度快；主题2中的高频特征词有“正品”、“不错”、“发货”、“速度”、“收到”、“很快”等，主要反映京东上的iPhoneX是正品，质量有保证。
- 根据对iPhoneX差评的3个潜在主题的特征词提取，主题0中的高频特征词有“发现”、“贵”、“客服”、“重启”等，主要反映iPhoneX贵、客服服务不好等；主题1中的高频特征词有“舒服”、“毛病”、“屏幕”等，主要反映iPhoneX存在一些毛病，尤其屏幕；主题2中的高频特征词有“慢”、“发现”、“找”、“真的”、“充电”、“充电器”等，主要反映iPhoneX充电器充电慢。