目录

[一、引言 2](#_Toc71240203)

[1.1 背景 2](#_Toc71240204)

[1.2 意义 2](#_Toc71240205)

[1.3 问题描述 3](#_Toc71240206)

[二、景区及酒店印象分析 3](#_Toc71240207)

[2.1 数据准备 3](#_Toc71240208)

[2.1.1 数据描述 3](#_Toc71240209)

[2.1.2 数据预处理 5](#_Toc71240210)

[2.1.3 分词技术 6](#_Toc71240211)

[2.1.4 删除停用词 6](#_Toc71240212)

[2.1.5 数据准备结果 7](#_Toc71240213)

[2.2 提取热门词 7](#_Toc71240214)

[2.2.1 热度词定义 7](#_Toc71240215)

[2.2.2 TD-IDF介绍 7](#_Toc71240216)

[2.2.3 热度词提取 8](#_Toc71240217)

[三、景区及酒店的综合评价 9](#_Toc71240218)

[3.1 数据准备 9](#_Toc71240219)

[3.1.1 数据描述 9](#_Toc71240220)

[3.1.2 数据预处理 10](#_Toc71240221)

[3.2 特征提取 10](#_Toc71240222)

[3.3 命名实体识别 12](#_Toc71240223)

[3.3.1 CRF原理 12](#_Toc71240224)

[3.3.2 BiLSTM原理 13](#_Toc71240225)

[3.4 实体情感识别 14](#_Toc71240226)

[3.4.1 实体分类 14](#_Toc71240227)

[3.4.2 实体情感分类 15](#_Toc71240228)

[3.5 建立得分体系 15](#_Toc71240229)

[四、网评文本的有效性分析 16](#_Toc71240230)

[4.1 问题分析 16](#_Toc71240231)

[4.1.1 重复或相似评论 16](#_Toc71240232)

[4.1.2 垃圾评论 16](#_Toc71240233)

[4.1.3 无意义评论 16](#_Toc71240234)

[4.2 数据准备 16](#_Toc71240235)

[4.2.1 数据描述 16](#_Toc71240236)

[4.2.2 数据预处理 17](#_Toc71240237)

[4.3 重复或相似评论的识别 17](#_Toc71240238)

[4.4 垃圾评论 19](#_Toc71240239)

[4.4.1 朴素贝叶斯分类模型 19](#_Toc71240240)

[4.4.2 应用原理 19](#_Toc71240241)

[五、景区及酒店的特色分析 20](#_Toc71240242)

[5.1. 数据准备 20](#_Toc71240243)

[5.1.1 数据描述 20](#_Toc71240244)

[5.1.2 数据预处理 21](#_Toc71240245)

[5.2 LDA主题模型 22](#_Toc71240246)

[5.2.1 LDA模型介绍 22](#_Toc71240247)

[5.2.2 LDA核心思想 22](#_Toc71240248)

[5.2.3 模型实现 22](#_Toc71240249)

[5.3 结果分析 22](#_Toc71240250)

[六、总结 23](#_Toc71240251)

# 一、引言

## 1.1 背景

近日，中国互联网络信息中心（CNNIC）在京发布第47次《中国互联网络发展状况统计报告》（以下简称《报告》）。《报告》显示，截至2020年12月，我国网民规模达9.89亿，较2020年3月增长8540万，互联网普及率达70.4%。

在虚拟世界还不能代替旅游业的时代，假期中的旅游行业依然是非常火爆。由于旅游APP的发展，人们开始通过APP来安排自己的假期旅游活动。曾经景区和酒店都会设置留言本来记录游客对该目的地的评价，而现在人们有了更方便的写评论的方式——网络评价。人们在APP或者网站上对景区或者酒店进行体验后的评论，既方便了游客，又方便了管理人员。

随着旅游APP的普及，对大量复杂的游客评论进行快速处理变成急需解决的问题。将大数据技术引入旅游业，优化旅游业的服务、管理和监督效率已经是大势所趋。而基于大数据技术中的自然语言处理的游客评论分析可以避免“人工评论处理的效率低”这一问题，通过机器处理游客评论，提高旅游业的管理效率。因此，本文通过对景区和酒店的评论进行挖掘，提取目的地TOP20热门词，形成景区及酒店印象分析；结合景区及酒店的综合评价，对网评文本进行有效性分析，并建立模型分析景区及酒店的特色。

## 1.2 意义

提升景区及酒店等旅游目的地美誉度是各地文化旅游主管部门和旅游相关企业非常重视和关注的工作，涉及到如何稳定客源、取得竞争优势、吸引游客到访消费等重要事项。游客满意度与目的地美誉度紧密相关，游客满意度越高，目的地美誉度就越大。因此掌握目的地游客满意度的影响因素，切实提高游客满意度、最终提升目的地美誉度，不仅能够保证客源稳定，而且对于旅游企业科学监管、资源优化配置以及市场持续开拓具有长远而积极的作用。

## 1.3 问题描述

（1）任务一：在分析游客对景区或者酒店的评价情况时，如果查看原数据中毫无规范的评论，不仅浪费时间而且效率很低。如果把评论提取关键词，能够快速了解评论中景区或酒店的大概描述，这样不仅提高了分析效率，并且一眼就能看出目的地的特点。于是依据附件中景区及酒店网评文本，按赛题表1格式计算出目的地TOP20热门词，并保存为文件“印象词云表.xls”。

（2）任务二：景区及酒店的综合评价，根据附件1景区及酒店网评文本及附件2景区及酒店得分建立合理的数学模型及相应算法，按满分为5分对景区及酒店的服务、位置、设施、卫生、性价比五个方面进行评分，并按照均方误差（Mean Squared Error，MSE）进行模型评价。

（3）任务三：网评文本的有效性分析，处于各种原因，网络评论常常出现内容不相关、简单复制修改和无有效内容等现象，妨碍了游客从网络评论中获得有价值的信息，也为各网络平台的运营工作带来了挑战。请从文本分析的角度，建立合理的模型，对附件1景区及酒店网络评论的有效性进行分析。

（4）任务四：景区及酒店的特色分析，旅游业繁荣发展给游客带来了选择困难的问题，评分接近的景区或酒店很难根据评分进行取舍。建立合理的模型和算法，从景区及酒店的网评文本中挖掘他们各自的特色和亮点，以吸引游客提升竞争优势。请选择综合评价高、中、低三个层次的各3家景点和3家酒店，结合模型的结果，分析他们各自的特色。

# 二、景区及酒店印象分析

## 2.1 数据准备

### 2.1.1 数据描述

本题需要两个数据文件，一个是“景点评论.xlsx”和“酒店评论.xlsx”。

“景点评论.xlsx”一共有59106条数据。每一条数据由“景区名称”，“评论日期”，“评论内容”组成。评论内容包括中文和英文。图2.1.1-1显示“景点评论.xlsx”样式，图2.1.1-1显示了景点名称及其评论数占比情况。

“酒店评论.xlsx”一共25225条数据。每一条数据由“酒店名称”，“评论日期”，“评论内容”，“入住房型”组成。评论内容包括中文和英文。图2.1.1-3显示“酒店评论.xlsx”样式，图2.1.1-4显示酒店名称及其评论数占比情况。

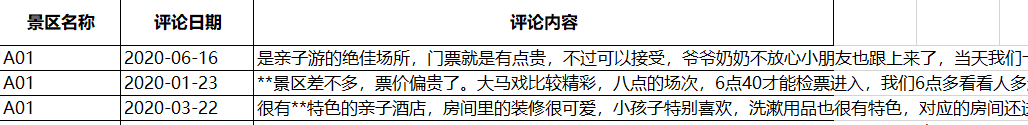


图2.1.1-1“景点评论.xlsx”样式

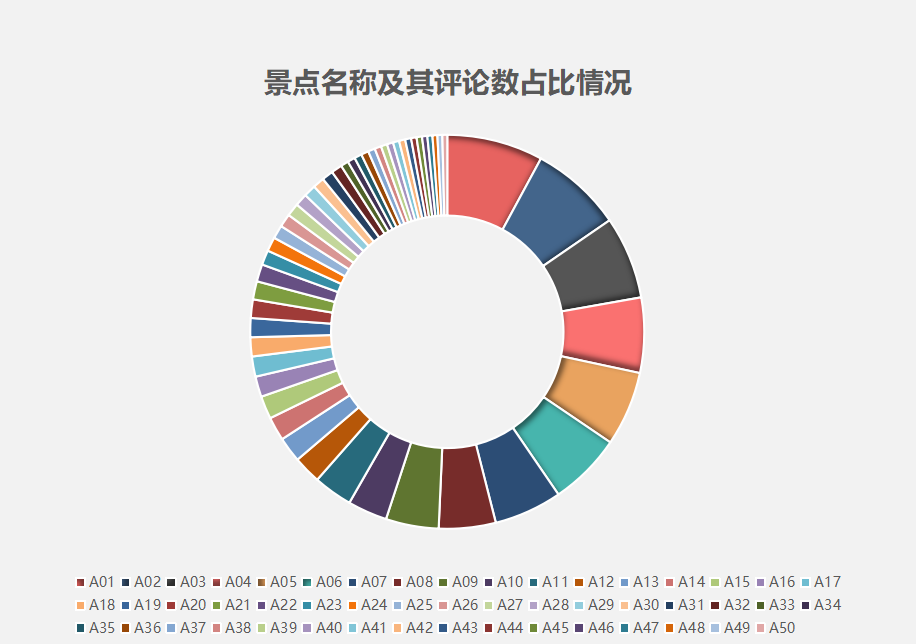


图2.1.1-2景点名称及其评论数占比情况

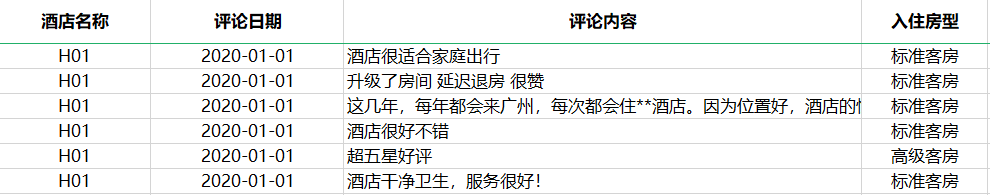


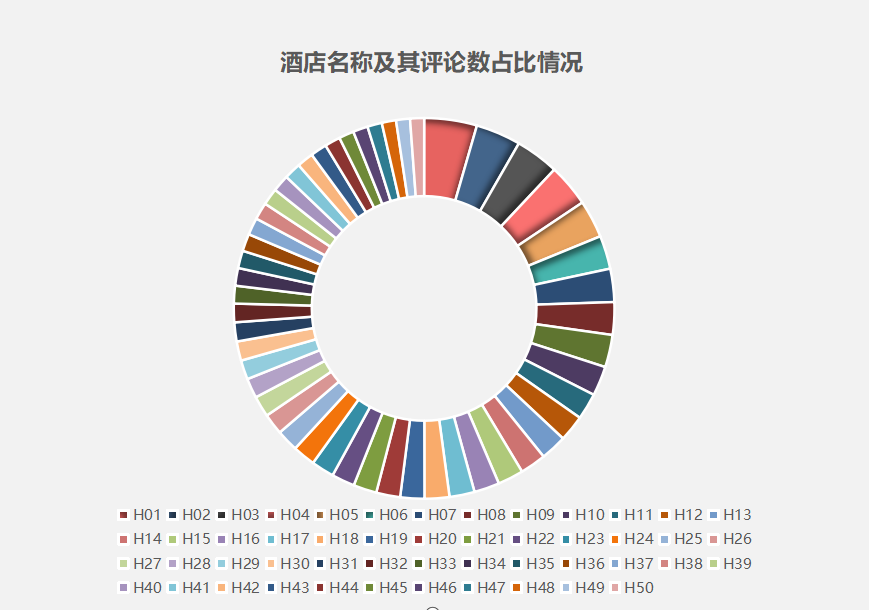
图2.1.1-3“酒店评论.xlsx”样式

图2.1.1-4酒店名称及其评论数占比情况

### 2.1.2 数据预处理

任务1的评论提取热度词之前，首先要对数据集进行数据清洗。由于评论语句随意性和可能的其他因素导致评论乱码问题会使得评论文本不规范。在给定的数据文件中主要是对“评论内容”进行数据预处理。

1.去掉评论时间

（1）原因：任务一是计算出目的地TOP20热门词，和评论时间没有太大关系，所 以可以不考虑评论时间。

（2）处理方法：在之后产生新的数据文件时不导入其中。

2.去除数字

（1）原因：在浏览评论数据发现，大部分数字都是关于景区或者酒店名称，或者是酒店的房间号，服务员的工号等一些无意义的数字，属于无价值信息。虽然停用词表中有数字，但是由于分词的结果数字无法和停用词中的数字匹配，所以需要在预处理阶段先把评论中的数字去掉。

（2）处理方法：在Python中利用正则表达式去掉原始数据集中的数字。

3.去重复评论

（1）原因：在浏览了评论数据发现，有些评论对应的景区名称和评论内容一模一样，评论对应的酒店名称和入住房型、评论内容一模一样，但是评论时间不一样。造成的原因有很多，但是其原因带来的结果就是重复的评论没有参考意义。而由于题目要求酒店评论不考虑酒店房型问题，所以同样只考虑同一酒店名称和评论内容。

（2）处理方法：在Python中利用函数去掉重复数据。

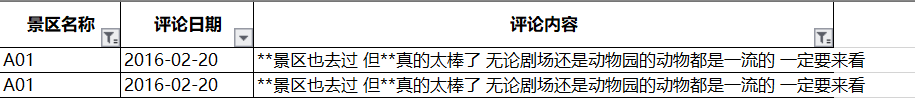


图2.1.2-1 “景区评论.xlsx”重复数据示例

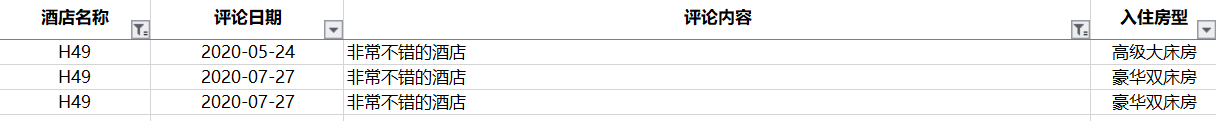


图2.1.2-2 “酒店评论.xlsx”重复数据示例

4.形成新的评论数据文件

（1）原因：因为评论日期不在考虑范围内，生成新的数据文件以目的地为核心，把同一目的地的评论集合在一起，即新的数据文件中每一行包括目的地和处理后的该目的地所以的评论集合。形成新的文件方便后续热度词提取。

（2）处理方法：在Python中将需要的数据进行以上处理之后导入新的csv文件中。

（3）图2.1.2-1 显示了景点评论新文件，图2.1.2-2 显示酒店评论新文件。由于题目要求，本文在处理酒店评论时将房型排除在外不参与考虑中。

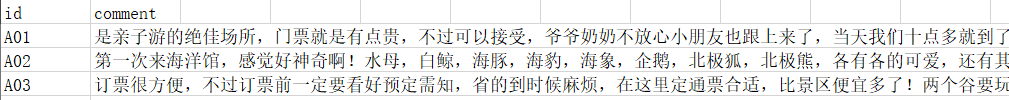


图2.1.2-1 景点评论新文件

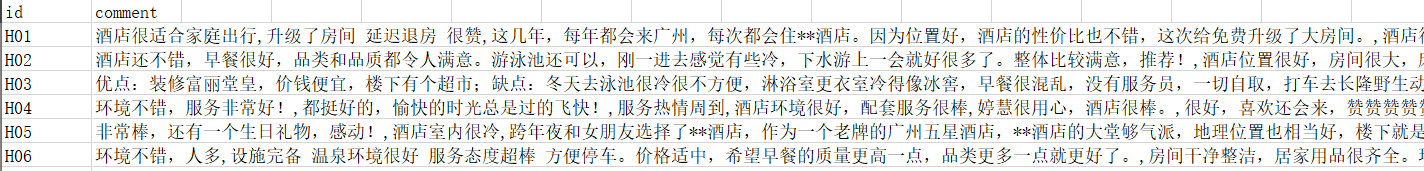


图2.1.2-2 酒店评论新文件

### 2.1.3 分词技术

中文分词的原理是将一个汉字序列进行切分，从而得到一个个独立的词。在中文文本中，句子和段落之间利用标点符号分隔，但是每一个词之间没有分隔号。所以首先需要在处理的文本中分隔出独立的词。

英文分词和中文分词不太一样，英文分词一个单词就是一个词，词之间有明显的区分。

目前中文分词有Jieba分词工具，pkuseg分词工具和thucla分词工具。

（1）Jieba分词工具：专用于分词的 Python 库。支持三种分词模式：精确模式，试图将句子最精确地切开，适合文本分析。全模式，将句子中所有的可能成词的词语都扫描出来，速度快，但是不能解决歧义。搜索引擎模式：在精确模式的基础上，对长词再次切分，提高召回率。

（2）pkuseg分词工具：多领域中文分词工具，此工具包为不同领域的数据提供个性化的预训练的模型。根据待分词文本的领域特点，使用者可以选择不同的模型。目前提供旅游领域模型适合本文。

（3）thucla分词工具：由清华大学研制推出的一套中文词法分析工具包，具有中文分词和词性标注功能。具有以下特点：执行能力强，分词速度快，分词效率高。

（4）以上三种分词工具都能同时对中英文进行分词，但是由于pkuseg引入模型之后再对景区评论分词效果更好，所以对于景区评论使用的是pkuseg分词，对酒店评论使用的是thucla分词。

### 2.1.4 删除停用词

停用词是指那些在文本中大量出现但是没有什么参考价值的词语。比如“吧”“啊”“的”“且”等。由于在数据预处理时没有对一些标点符号，特殊符号等进行处理，还有一些常规英文单词需要过滤。

以上都为比较常规的停用词，对于当前文本任务还需要过滤一些特定的停用词，比如“酒店”，“景区”等。

本文使用的停用词表示例如下图2.1.4-1

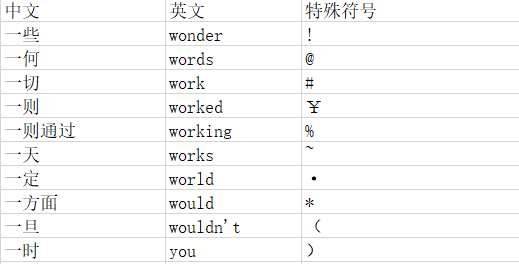


图2.1.4-1 停用词表示例

### 2.1.5 数据准备结果

经过上述的预处理，分词和删除停用词之后，得到较为精确的分词结果。

## 2.2 提取热门词

### 2.2.1 热度词定义

本文认为的热度词为“评论中该次出现次数多的词”。如果评论中提及次数太低并不足以成为热度词。

### 2.2.2 TD-IDF介绍

TF-IDF（词频-逆向文件频率）是一种用于信息检索与文本挖掘的常用加权技术。TF-IDF是一种统计方法，用来评估一个词语对于一个文本集中一个文本的重要程度。每个词的重要程度随着它在一个文本中出现的次数成正比增加，但也随着在文本集中出现的频率成反比下降。

TF-IDF主要思想：如果一个词在一个文本中的词频高，并且在其他文本中出现次数比较少，那么这个词比较适合用来分类。

在这个旅游景点和酒店遍地开花的时代，无论是景区还是酒店，对于游客来说，知道景区或者酒店的特色固然重要，但是更重要的是知道这个景区或者酒店相比其他的景区和酒店更胜一筹的地方。而TD-IDF算法在一定程度上契合本文的想法。

（1）TF（词频）

TF表示一个词在文本中出现的频率。

（2）IDF（逆向文件频率）

某一特定词语的IDF，可以由总的文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。如果包含词条t的文档越少, IDF越大，则说明词条具有很好的类别区分能力。（是防止词不在文本集中使得分母为0）

（3）

结合目的文本中的高词频率和该词在文本集中的低文本频率，可以产生出高权重的 TF-IDF。因此，TF-IDF倾向过滤掉常见的词语，留下重要程度较大的词。

### 2.2.3 热度词提取

用TF-IDF算法对景区评论和酒店评论进行热度词提取，并将计算的热度显示出来，保存在csv文件中。下列图为提取结果：

景区：图2.2.3-1Jieba + TF-IDF示例，图2.2.3-2 thucla + TF-IDF示例，图2.2.3-3 pkuseg + TF-IDF示例

图2.2.3-1 Jieba+TF-IDF示例图2.2.3-2 thucla+TF-IDF示例 图2.2.3-3 pkuseg+TF-IDF示例

酒店：图2.2.3-4 Jieba + TF-IDF示例，图2.2.3-5 thucla + TF-IDF示例，图2.2.3-6 pkuseg + TF-IDF示例

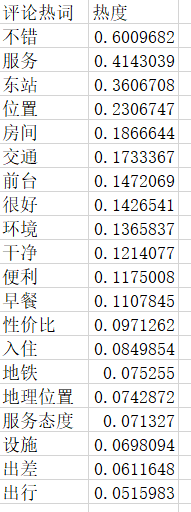
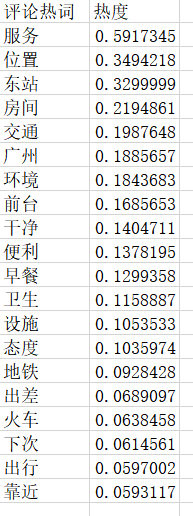
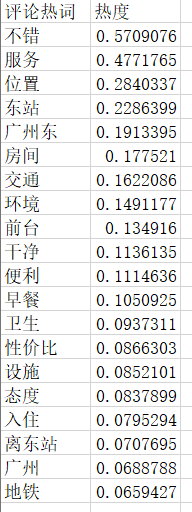
  

图2.2.3-4 Jieba+TF-IDF示例图2.2.3-5 thucla+TF-IDF示例 图2.2.3-6 pkuseg+TF-IDF示例

# 三、景区及酒店的综合评价

## 3.1 数据准备

### 3.1.1 数据描述

任务二的数据集中包括附件1的景区及酒店网评文本和附件2的景区及酒店得分，

附件2包括不同景区和不同酒店各个方面的得分，如下图所示:



图3.1.1-1 景区得分



图3.1.1-2 酒店得分

### 3.1.2 数据预处理

任务二的数据预处理同任务一，进行一般的预处理流程包括去除特殊符号、分词、去停用词等

## 3.2 特征提取

将文本输入分类器之前，需要将文本转化为计算机所识别的符号数学形式， NLP 中最直观的表示方式就是 One-hot Representaion，它将每个词表示为一个向量，长度为词表大小，只有一个维度值为 1，但是这种表示方式没有考虑语义，存在“词汇鸿沟”的缺陷。因此一般采用 Distributed Representaion，它表示低维实数向量，常见维度为 50 维和 100 维。这种方法可以使相关或相似的词在距离上更接近。常用的训练词向量方法有 Word2Vec 与FastText。

1. One-hot Representation

One-hot Representation 即用一个很长的向量来表示一个词，向量长度为词典的大小N，每个向量只有一个维度为1，表示该词语在词典的位置，其余全部为0。

举例:

“话筒”表示为 [0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 …]

“麦克”表示为 [0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 …]

这种 One-hot Representation 如果采用稀疏方式存储，会是非常简洁，也就是给每个词语分配一个数字 ID。但这种表示方式有两个缺点：

（1）容易受维数灾难的困扰，每个词语的维度就是语料库字典的长度。

（2）词语编码往往是随机的，导致不能刻画词与词之间的相似性

2. Distributed representation

Distributed representation 最早由 Hinton在1986 年提出。其依赖思想是：词语的语义是通过上下文信息来确定的，即相同语境出现的词，其语义也相近。

3. Distributed Representation与one-hot representation对比

在形式上，one-hot representation 词向量是一种稀疏词向量，其长度就是字典长度，而Distributed Representation是一种固定长度的稠密词向量。一般长这样：[0.792, −0.177, −0.107, 0.109, −0.542, …]

在功能上，Distributed representation 最大的贡献就是使相关或者相似的词，在距离上更接近了。

1. Word2vec介绍

Word2Vec 实际上是两种不同思想实现的：CBOW（Continuous Bag of Words) 和 Skip-gram。CBOW的目标是根据上下文来预测当前词语的概率，且上下文所有的词对当前词语出现概率的影响的权重是一样的，因此叫continuous bag-of-words模型。如在袋子中取词，取出数量足够的词就可以了，至于取出的先后顺序是无关紧要的。Skip-gram刚好相反：根据当前词语来预测上下文的概率。这两种方法都利用人工神经网络作为它们的分类算法。起初每个单词都是一个随机 N 维向量。经过训练之后，该算法利用 CBOW 或者 Skip-gram 的方法获得了每个单词的最优向量。训练过程如下图所示：

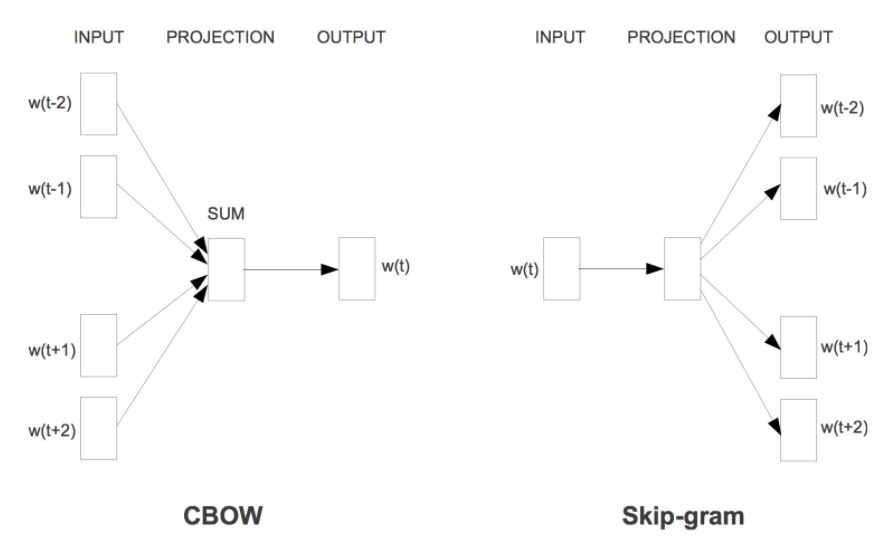


图3.2.4-1 word2vec两种不同模型的训练过程

## 3.3 命名实体识别

### 3.3.1CRF原理

条件随机场（Conditional Random Field，CRF）是自然语言处理的基础模型，广泛应用于中文分词、命名实体识别、词性标注等标注场景。条件随机场CRF与深度学习结合，产生了BiLSTM-CRF、BiLSTM-CNN-CRF等模型，在中文分词、命名实体识别、词性标注也取得不错的效果。CRF层可以加入一些约束来保证最终预测结果是有效的。这些约束可以在训练数据时被CRF层自动学习得到。有了这些有用的约束，错误的预测序列将会大大减少。

crf中的损失函数包括两种类型的分数：发射分数和转移分数，分别由状态特征函数和转移特征函数计算得到。本文中使用的的是BILSTM-CRF模型进行命名实体识别，因此发射分数即为BILSTM层的输出。而转移分数可以由CRF层自行学习。比如下图中使用CRF进行词性标注，发射分数s[x3][art.]是单词x3词性为art.的分数，转移分数t[v.][art.]即是词性v.到词性art.的分数。crf层便使用这两种分数组成损失函数，从而将问题转化成了一个最优化问题。

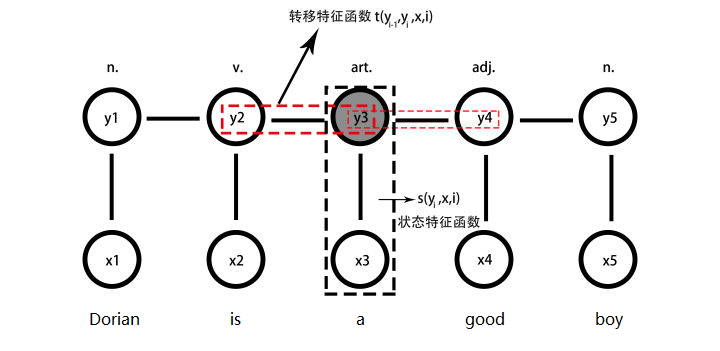


图3.3.1-1 条件随机场的原理

### 3.3.2 BiLSTM原理

LSTM的全称是Long Short-Term Memory，它是RNN（Recurrent Neural Network）的一种。LSTM由于其设计的特点，非常适合用于对时序数据的建模，如文本数据。BiLSTM是Bi-directional Long Short-Term Memory的缩写，是由前向LSTM与后向LSTM组合而成。两者在自然语言处理任务中都常被用来建模上下文信息。

下图是lstm总体框架。LSTM模型是由时刻的输入词x,细胞状态C,临时细胞状态,隐层状态h,遗忘门f,记忆门i,输出门o组成。LSTM的计算过程可以概括为，通过对细胞状态中信息遗忘和记忆新的信息使得对后续时刻计算有用的信息得以传递，而无用的信息被丢弃，并在每个时间步都会输出隐层状态h，其中遗忘，记忆与输出由通过上个时刻的隐层状态h和当前输入x计算出来的遗忘门f，记忆门i，输出门o来控制。

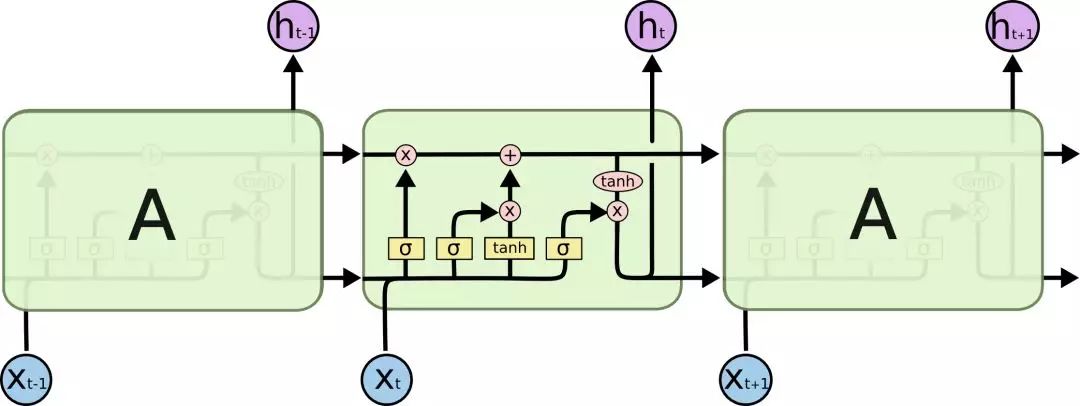


图3.3.2-1 LSTM框架

前向的LSTM与后向的LSTM结合成BiLSTM。因为LSTM中每个细胞每个细胞只会受之前细胞的隐层状态h和当前输入x影响，为了根据上下文建模，需要使用BILSTM，使其包含前向与后向的所有信息

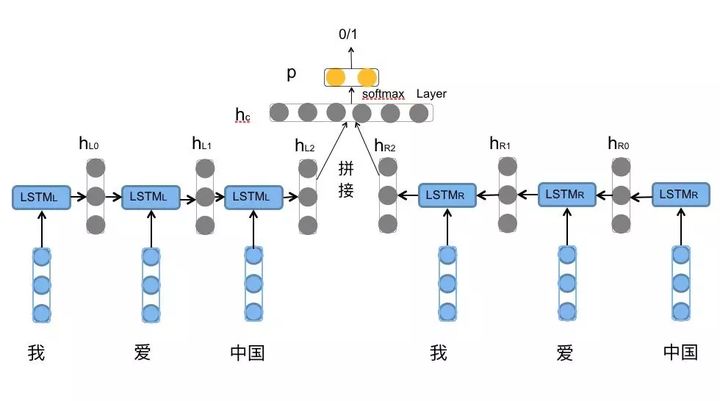


图 3.3.2-2 BiLSTM框架

## 3.4 实体情感识别

### 3.4.1 实体分类

将上一步3.3命名体识别后的实体进行分类，归类到我们需要的五个方面（服务、位置、设施、卫生和性价比）,如图3.4.1所示。这样就完成了不同方面的细分评价指标，使得能够对各个方面进行分别评价。

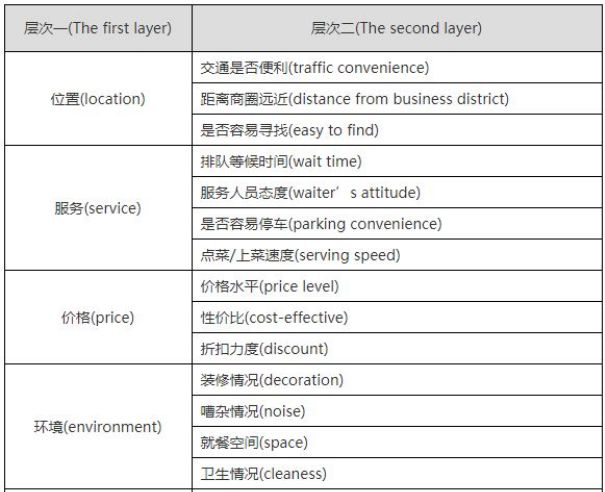


图3.4.1 实体归类

### 3.4.2 实体情感分类

通过大连理工情感词典识别评论中的情感词，根据情感极性保留其中的积极与消极倾向的情感词，同时记录情感词在评论中的位置。同时识别和、记录否定词和程度副词在评论中的位置。

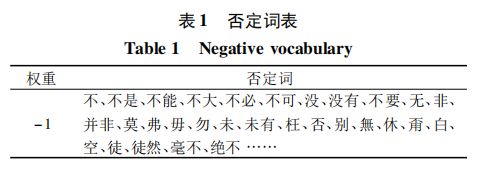


图3.4.2-1 否定词表

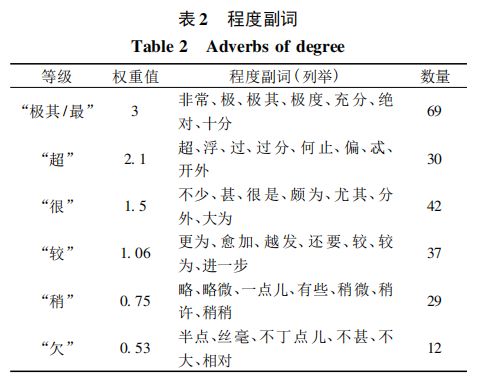


图3.4.2-2 程度副词词表

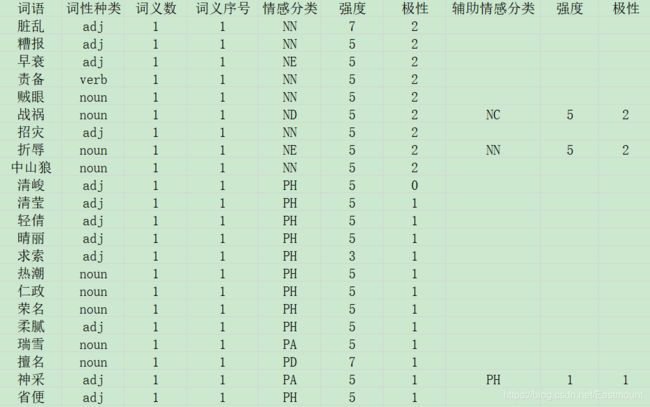


图3.4.2-3 情感词表

## 3.5 建立得分体系

在获得了五个方面不同实体的情感倾向之后，就可以对其进行有效的评分操作，建立合理的打分机制来计算得分。首先对于之前处理过的情感词，将其情感极性进行转化，计算出相应的情感值=极性（处理后）\*强度。然后向前查找已识别的命名实体，若遇到否定词/程度副词则根据对应权重进行情感值处理，最后将情感值赋予第一个查找到的命名实体，通过实体所属分类将情感值以一定权重赋予对应的方面。

# 四、网评文本的有效性分析

## 4.1 问题分析

### 4.1.1 重复或相似评论

在大量的网评文本当中，存在着许多重复或者相似度极高的评论，由于这些评论中有着类似的内容，不利于游客对景区及酒店好坏的判断，并且重复的评论内容会给服务器增加不必要的负担，我们需要识别出这些重复或相似的评论以此判断出评论的有效性。



图4.1.1-1 重复的评论

### 4.1.2 垃圾评论

网评文本存在着许多的垃圾评论，评论所述并不围绕着评论属性，或者评论中存在着大量的低俗、暴力等不利于国家网络安全事业发展的词汇，这时候就需要我们有识别出这些垃圾文本内容的能力。此类任务目的在于识别出重复或相似评论以及垃圾评论（spam），垃圾文本内容识别问题是安全内容的一个重要方面，从类别上讲属于文本分析中的文本二值分类问题，即识别出评论是否为垃圾评论。

### 4.1.3 无意义评论

文本评论当中有些尽管看上去写了很多，但是和主题无关，这时候需要我们判断评论是否有效可行，符合评论主题。类似于下图的评论就被判断段无意义（原因是没有评论主体）：



图4.1.3-1 无意义评论

## 4.2 数据准备

### 4.2.1 数据描述

任务三的数据集和任务一相同，即附件1的景区及酒店网评文本。

### 4.2.2 数据预处理

该任务的数据预处理部分同上述任务二的预处理，即进行一般的去特殊字符（注意这时和任务二有所区别的地方在于，任务三不能去除网址链接等可以识别出垃圾评论的特征点）、去停用词、分词等处理。

## 4.3 重复或相似评论的识别

采用Simhash算法，将文本隐射成指纹，通过对比指纹来识别文本相似度。传统的对比文相似度的方法，大多是将文本分词之后，转化为特征向量距离的度量，比如欧式距离、海明距离等，但是缺点是无法适用于海量数据，因为维度过大而造成计算所耗资源过大。

Simhash算法的具体过程：  
 1.分词

对待处理文档进行中文分词，得到有效的特征及其权重。可以使用TF-IDF方法获取一篇文章权重最高的前topK个词（feature）和权重（weight）。即可使用jieba.analyse.extract\_tags()来实现

2.hash

对获取的词（feature），进行普通的哈希操作，计算hash值，这样就得到一个长度为n位的二进制，得到（hash：weight）的集合。

3.加权

在获取的hash值的基础上，根据对应的weight值进行加权，即W=hash\*weight。即hash为1则和weight正相乘，为0则和weight负相乘。例如一个词语经过hash后得到（010111：5）经过步骤（3）之后可以得到列表[-5,5,-5,5,5,5]。

4.合并

将上述得到的各个向量的加权结果进行求和，变成只有一个序列串。如[-5,5,-5,5,5,5]、[-3,-3,-3,3,-3,3]、[1,-1,-1,1,1,1]进行列向累加得到[-7，1，-9，9，3，9]，这样，我们对一个文档得到，一个长度为64的列表。

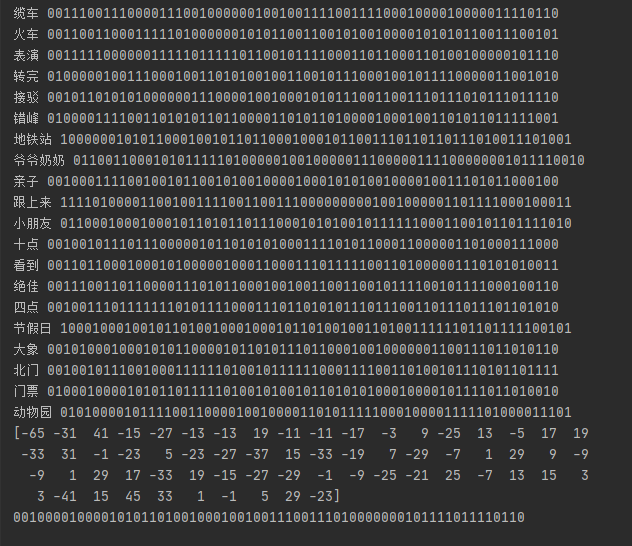
5.降维

对于得到的n-bit签名的累加结果的每个值进行判断，大于0则置为1，否则置为0，从而得到该语句的simhash值。例如，[-7，1，-9，9，3，9]得到 010111，这样，我们就得到一个文档的simhash值。

最后根据不同语句的simhash值的汉明距离来判断相似度。



图 4.3-1 A01景区某评论

图 4.3-2 上述评论simhash值的计算过程

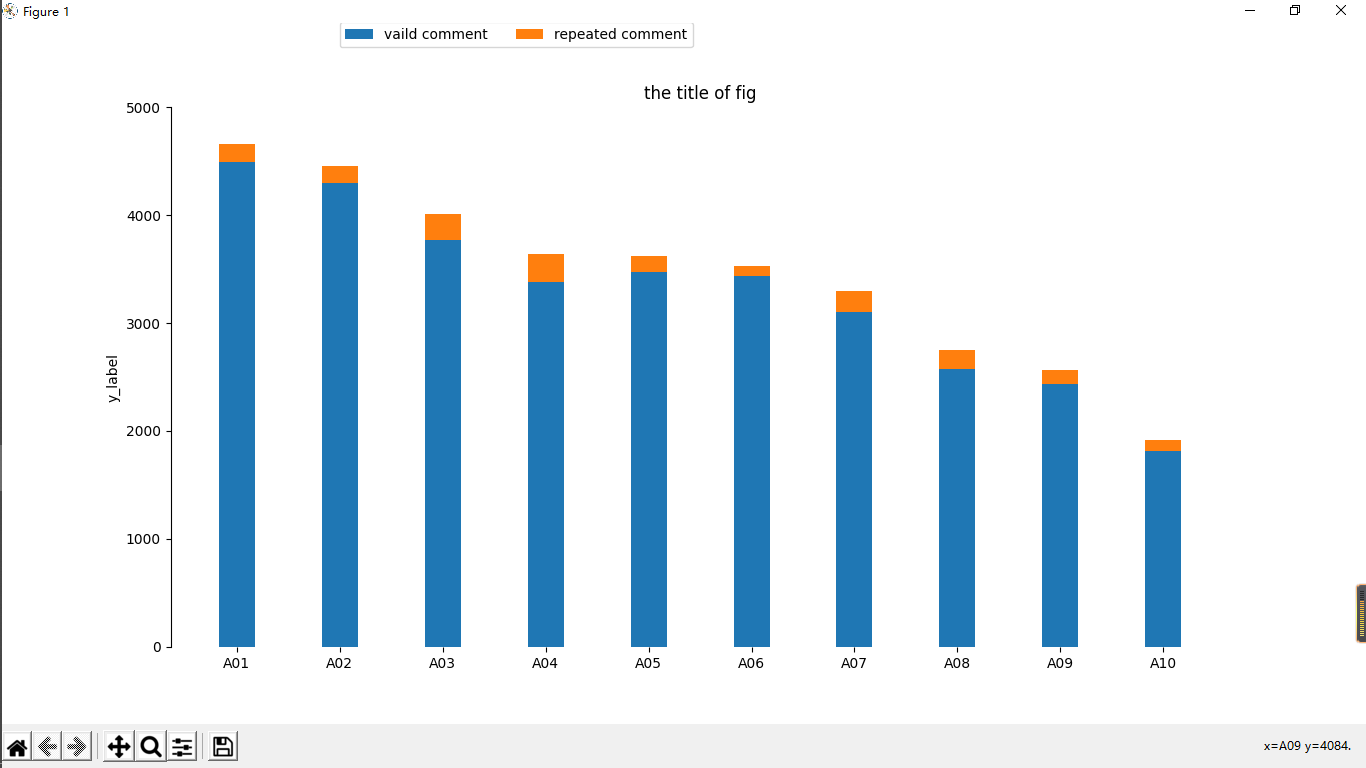


图4.3-3 网评文本当中重复或相同评论所占比例

## 4.4 垃圾评论

### 4.4.1 朴素贝叶斯分类模型

朴素贝叶斯算法是被工业界广泛应用的机器学习算法，它有较强的数学理论基础，在一些典型的应用中效果显著。朴素贝叶斯算法基于概率论的贝叶斯理论。该理论的核心公式如下：



式中，c表示某种分类，则表示已知X的情况下，类型为c的条件概率。我们求出各个类别下的，然后比较它们的大小，以概率最大的作为最后的类别，以此达到目的。

计算条件概率的方法：已知，则。朴素贝叶斯假定互为独立变量，则

而（I为指数函数，存在则为1，不存在则为0），和都可以用训练数据直接统计得出。故依据上述分析求得的大小。又由于对所有类别都是固定大小，所以比较条件概率的大小等同于比较的大小，这就是朴素贝叶斯的数学定理。

### 4.4.2 应用原理

第一步：对于一个评论来说，它由多个单词组成，在考虑评论是否为垃圾评论的时候，先简化问题为考虑包含某一个特定的单词的评论是否为垃圾评论：



上式计算了在已知某个单词的情况下这个评论是spam的可能性，这个概率由先验概率进行计算，式中

* 代表一个评论是spam的概率
* 代表一个评论是ham的概率
* 代表了在spam评论中这个单词出现的概率
* 代表了在ham评论中这个单词出现的概率

第二步：在得到评论中出现某一个单词它的垃圾评论的可能性后，要考虑的就是这个评论含的所有的单词对于这个评论是垃圾评论可能性的影响。如果按概率论来讨论所有的单词出现在这个评论中的概率会十分复杂，这里就假设所有的单词出现在这个评论中是条件独立的事件，那么对于已知评论中所有的单词，求这条评论为垃圾评论的概率的公式可以简化为：



式中，，

第三步：在得到概率之后，和我们预定的阈值进行比较，如果大于阈值，就可以判断为垃圾评论。

# 五、景区及酒店的特色分析

## 5.1. 数据准备

### 5.1.1 数据描述

针对任务四的数据有：总得分高、中、低三个层次的各3家景点和3家酒店的所有的“评论内容”和相应的“评论时间”。“评论时间”是评论发表的具体时间，“评论内容”是评论的详细内容。

根据酒店和景点的评分表，选取符合条件的9家酒店和9家景点。

高分酒店：H06、H07、H13

中等酒店：H09、H35、H47

底分酒店：H38、H39、H43

高分景区：A23、A38、A39

中等酒店：A12、A25、A46

低分酒店：A02、A04、A05

### 5.1.2 数据预处理

1. 数据去数字

评论去除数字和任务一的原因和步骤一致，这里不再赘述。

1. 数据去重

为了保证分析结果的准确性，评论去重比任务一的步骤要更为精确一些。除了一模一样的评论需要过滤掉以外，本文使用基于simhash算法去重，可以去除一些相近的表达。

1. simhash算法
2. 算法介绍

simhash是一种计算文本相似度的hash算法。通过simhash能将一个文本映射成64bit，再比较两个文本的64bit的海明距离得到的相似程度。

1. 核心思想

将高维的特征向量映射成低维的特征向量（将文档降维到hash数字），通过两个向量的海明距离来确定文本是否重复或高度近似。

1. 算法步骤：分词、hash、加权、合并、降维。图

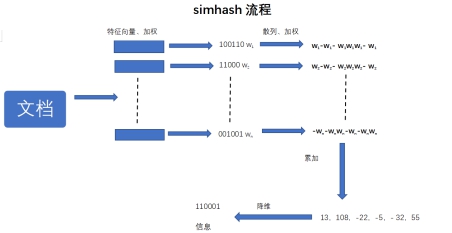


图5.1.2-3-1 simhash原理图

1. 数据归集

（1）归集原因：游客对酒店和景区的评论还需要根据时间来进行归集，本文将年月份和目的地作为归集元素。

（2）处理方法：用Python处理。（图5.1.2-4-1为原始数据集将年月提取之后示例）

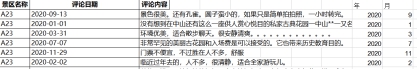


图5.1.2-4-1原始数据集将年月提取示例

1. 分词和去停用词同样与任务一的处理一致，这里不再赘述。
2. 对评论数据进行分析

进行情感倾向性分析，将评论数据分割成正面、负面、中性。抽取正面和负面两组进行词义网络的构建与分析。

## 5.2 LDA主题模型

本文针对已有的酒店和景区文本评论数据进行建模分析。进行基本的预处理、分词等操作后建立LDA主题模型，实现对文本评论数据的倾向性判断及信息挖掘分析。

### 5.2.1 LDA模型介绍

LDA模型采用词袋模型将每一篇文档视为一个词频向量，从而将文本信息转化为易于模的数字信息。

### 5.2.2 LDA核心思想

LDA主题的核心思想是：一个文本以一定的概率选择了某个主题，并从这个主题中以一定概率选择了某个词。



### 5.2.3 模型实现

1. 将每个酒店或景区评论文本一份为二，分为正面评价和负面评价两个文本分析。

2. 对分类后的文本进行分词，分词后建立LDA模型。

3. 通过LDA主题分析，评论将聚成几个主题，每个主题下生成N个最有可能出现的词语以及相应的概率。

## 5.3 结果分析

根据对三个等级共9家酒店和三个等级9个景区评论进行主题分析，可以提取到每一家酒店或者每一个景区的“亮点”。

# 六、结语

由于互联网的普及，数据爆发式增长，如何处理数据并从数据中获取有价值的信息成为了世界的焦点。互联网交叉旅游领域，为了满足大量游客评论文本数据的挖掘和分析要求，本文基于细粒度情感分析、simhash算法、朴素贝叶斯分类器、LDA主题模型等将游客对酒店和景区的评论和评分进行分析并解决四个问题。在模型的设计方面虽然存在着许多不足，但是通过这次竞赛也让我们学习到了许多NLP相关知识，了解了机器学习以及深度学习的一些算法和模型，真正意义上拓宽了我们的视野。