基于自然语言处理的茂名旅游知识图谱构建与分析

摘要

互联网的发展使得旅游资源与相关数据在网络上能够很容易被获取到，但由于相关数据体量庞大，来源错综复杂，模态形式繁多，因此不易于处理。对于2018-2021四年茂名市旅游的相关OTA、UGC数据，综合运用多种NLP技术手段进行了一定探究。

对于问题一，将其抽象为**文本二分类**问题，通过正则表达式获得文本标注以后对比了TextCNN、TextRNN、TextRCNN、HAN、Transformer、DPCNN、BERT和ERNIE八种深度学习文本分类器的效果，发现在对于此类文本分类问题上TextCNN不仅速度更快并且分类准确率也是最高的，注意力类模型在此数据集上的表现反而较差，这说明并非越复杂的模型效果越好。

对于问题二，需要综合运用多种模型。产品提取本质上是一个**命名实体识别**问题，先利用开源TexSmart模型对问题进行初步标注以后对比了HMM、CRF、BiLSTM、BiLSTM-CRF和BERT-BiLSTM-CRF五种模型，发现BiLSTM-CRF模型的效果最佳。在成功识别产品以后我们选择环境、服务、产品质量和成本四项维度，并将情绪作为调节变量，构建TOPSIS综合**评价模型**得到热度评分，并分析了情绪变量在整体热度评价中的调节效应，最终构建了不同维度之间的结构方程模型揭示变量作用机理。

对于问题三，将其抽象为**关联关系挖掘**问题，使用Apriori算法和改进的PCY算法抽取了频繁项集和关联关系模式，并通过统计方法综合发现了产品大类之间的关联模式强度。

对于问题四，选择通过Neo4j构建**图数据库**存储产品实体对象并构建层次关系三元组和关联关系三元组，从而形成茂名市本地旅游**知识图谱**，通过前端方法对**知识图谱**进行可视化。最后综合上述问题的结论书写建议信见正文。

模型结合了统计学与机器学习方法的优点，具有可靠性，并且结果具有良好的可解释性。综合了多种自然语言处理子任务和技术手段，灵活选择不同工具合理解决问题，融合数据驱动和知识驱动的双重视角，最终得到问题的合理解。从机器学习常用的衡量指标来看，模型的表现无疑是非常优秀的，能够高效率地进行茂名市本地旅游文本信息的挖掘分析和评价，并获得海量稳定安全的**图数据库**与**知识图谱**，在实际商业应用中有一定价值。

关键词：文本分类，命名实体识别，评价模型，关联关系挖掘，图数据库，知识图谱

Abstract

The development of Internet has made massive tourist information and relevant resources readily available. Nevertheless, due to the wide range of data resources, with its very multi-faceted modalities, it has brought us great difficulty in dealing with enormous data.

Generally speaking, aimed at the OTA and UGC data related to Maoming tourism within the four-year-period from 2018 to 2021, we employ NLP methods in a gesture to carry out a comprehensive exploration.

For problem 1, a binary-classification model is extracted. In data pre-procession, we first gained text annotations through regular expressions. Then we made comparisons of effects of **eight** deep learning **text classifiers** and discovered that TextCNN not only gains an advantage in speed but also boasts highest accuracy among all the rest. In contrast, the attention model's performance on this dataset shows less superiority in this case.

In regard to problem 2, a more integrated application of several models is required. After we had searched for references, we learned that product extraction is essentially a **Named Entity Recognition** problem. The open source TexSmart model was adopted by us to make the initial mark and then five models HMM, CRF, BiLSTM, BiLSTM-CRF and BERT-BiLSTM-CRF were compared, and the optimized result was found in BiLSTM-CRF model. Having successfully identified the product, we selected the four dimensions, environment, service, product quality and cost, and used emotion as a moderator variable to construct a TOPSIS comprehensive **evaluation model** and extracted a popularity score, afterwards we analyzed the moderating effect of emotional variables in overall popularity evaluation. A structural equation model between diverse dimensions was constructed in revealing the mechanism of variable action.

Problem 3 is boiled down to an association relationship mining problem. We first extract frequent itemsets and association relationship patterns by using the Apriori algorithm and the improved PCY algorithm, then we capture the strengths of the association patterns between the product categories with statistical methods involved.

In terms of problem 4, we set up a graph database through Neo4j to store product entity objects and constructed hierarchical relationship triples and association relationship triples, henceforth the local tourism knowledge map of Maoming City is formed, and we managed to visualize the knowledge map through front-end measures.

The conclusions of the four questions is summarized in the letter of recommendation.

In conclusion, our model combines the advantages of statistical and machine learning methods, is hence reliable, and the results are well interpretable. We comprehensively utilized a great range of natural language processing sub-tasks and technical approaches, flexibly selects different tools in determining reasonable resolutions, and integrates the dual perspectives of data-driven and knowledge-driven, and finally achieved an ideal result.

We efficiently conduct mining, analysis and evaluation of local tourism text information in Maoming City, and obtain a large number of stable and secure **graph databases** and **knowledge graphs**. Practically, the model contains certain value in real-life scenarios.

Keywords: text classification, named entity recognition, evaluation model, relationship mining, graph database, knowledge graph

目录

一. 问题背景与分析

1.1 问题重述

互联网的发展使得旅游资源与相关数据在网络上能够很容易被获取到。但由于相关数据体量庞大，来源错综复杂，模态多种多样，因此不易于处理。而在海量旅游有关文本中，用户与自媒体诞生的旅游数据（UGC）又与文本形式在线旅游文本（OTA）一同构成在线旅游数据的主要形式。OTA 和 UGC 数据的内容较为分散和碎片化，要使用它们对某一特定旅游目的地进行研究时，迫切需要一种能够从文本中抽取相关的旅游要素，并挖掘要素之间的相关性和隐含的高层概念的可视化分析工具。

本地旅游知识图谱是知识图谱的一种，在通用知识图谱的基础上加入了更多针对旅游行业的需求。它基于图数据库对知识进行存储，并采用图的形式直观全面地展示特定旅游目的地“吃住行娱购游”等旅游要素，以及它们之间的关联。利用知识驱动的方式结合自然语言处理（NLP）技术可以对本地旅游的不同维度、不同产品之间进行抽取与关系构建，从而将其直观化展示出来。

问题基于 2018-2020 年广东省茂名市的部分在线旅游数据，包括微信公众号文章、餐饮评论、酒店评论、游记、景区评论等不同评论对象数据，需要解决以下问题：

问题一：构建文本分类模型，对附件 1 提供的微信公众号的推送文章根据其内容与文旅的相关性分为“相关”和“不相关”两类，并将分类结果以表 1 的形式保存为文件“result1.csv”。与文旅相关性较强的主题有旅游、活动、节庆、特产、交通、酒店、景区、景点、文创、文化、乡村旅游、民宿、假日、假期、游客、采摘、赏花、春游、踏青、康养、公园、滨海游、度假、农家乐、剧本杀、旅行、徒步、工业旅游、线路、自驾游、团队游、攻略、游记、包车、玻璃栈道、游艇、高尔夫、温泉等等。

问题二：从附件提供的 OTA、UGC 数据中提取包括景区、酒店、网红景点、民宿、特色餐饮、乡村旅游、文创等旅游产品的实例和其他有用信息，将提取出的旅游产品和所依托的语料以表 2 的形式保存为文件“result2-1.csv”。建立旅游产品的多维度热度评价模型，对提取出的旅游产品按年度进行热度分析，并排名。将结果以表 3 的形式保存为文件“result2-2.csv”。

问题三：依据提供的 OTA、UGC 数据，对问题 2 中提取出的旅游产品进行关联分析，找出以景区、酒店、餐饮等为核心的强关联模式，结果以表 的形式保存为文件“result3.csv”。在此基础上构建本地旅游图谱并选择合适方法进行可视化分析。鼓励参赛队挖掘旅游产品间隐含的关联模式并进行解释。

问题四：基于历史数据，使用本地旅游图谱作为分析工具，分析新冠疫情前后茂名市旅游产品的变化，并撰写一封不超过 2 页的信件向该地区旅游主管部门提出旅游行业发展的政策建议。

1.2 问题分析

这一问题是一个自然语言处理与知识工程的综合性问题，需要综合运用多种模型与算法进行解决，并合理利用工具。经分析，我们认为这四项问题的大致路径如下：

任务一需要我们对微信公众号文章进行分类，判定是否与文旅相关。由于任务中已经提示了与文旅有关的文章相关的高频主题词，故可以基于关键词与正则表达式对文本进行初步标注后训练文本分类器。对于文本分类，这是一个经典的自然语言处理任务，基本思想为“上游嵌入表示+下游分类模型”。随着深度学习发展，基于神经网络的文本分类模型种类繁多。这里我们可以主要将其分为两个系列：不基于大规模预训练模型的分类和基于大规模预训练模型的分类。在建模中我们将对比 TextCNN、TextRNN、TextRCNN、DPCNN、HAN、Transformer、BERT、ERNIE 几类模型的效果。

任务二是一个复杂任务，需要综合运用多种模型。对于旅游产品的提取，这是一个命名实体识别问题，我们基于 TexSmart 对数据进行标注以后可以训练命名实体识别模型评估。而对于多维热度评价模型，我们可以有两类方案：我们可以从成本、产品质量、环境、服务四个角度抽取对应的子句进行评分，再进行整体文本的评分与情感分析，获得得分以后第一类方案是直接运用 TOPSIS 与熵权法进行综合评价；第二类方案是将每个语料中出现的实体拼接起来，类比 NLP 中的 TF-IDF 值计算出每个语料的 TF-IDF 值乘对应语料评分后规约化即可得到热度，再根据热度与不同维度之间的回归分析进行实证研究。

任务三是寻找不同实体之间的关联关系，我们基于任务二中将每个语料出现的实体拼接起来，可以运用 apriroi 算法或者引入哈希的改进算法 PCY 进行求解。除了可以得到两两配对的实体关联规则以外还可以得到更复杂的实体关联关系。

任务四构建知识图谱并分析，需要使用数据库技术。Neo4j 是一种典型的图数据库，它以三元组的形式存储知识，基于 Cypher 语言可以像 SQL 语句一样可以查询实体及其背后的关联关系。我们基于问题二和问题三的结果将不同实体和关联关系在数据库中构建节点及其关系后即可清晰分析两个知识图谱。

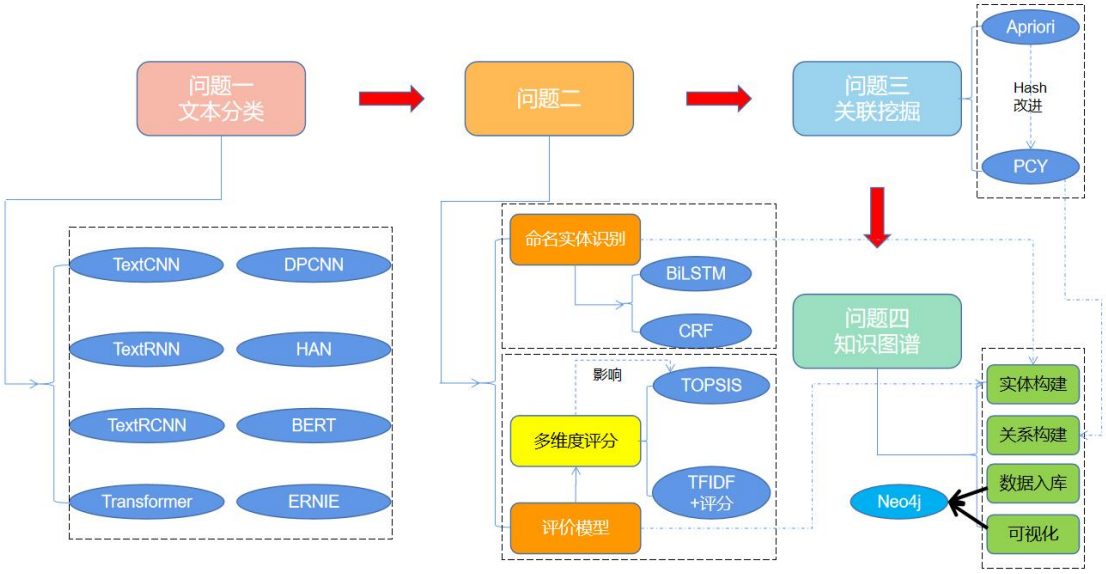


图 1. 问题的求解流程

二. 数据集描述与分析

2.1 数据集描述

这一数据集是 2018-2020 年茂名市旅游部分 UGC 和 OTA 数据，包括微信公众号文章、酒店评论、餐饮评论、景区评论和游记五个不同的数据。

对于微信公众号，其文章并不全是与文旅相关。我们对缺失内容按照 0 填充法进行空值填充以后将微信公众号的文本特征展示如图 2 所示：

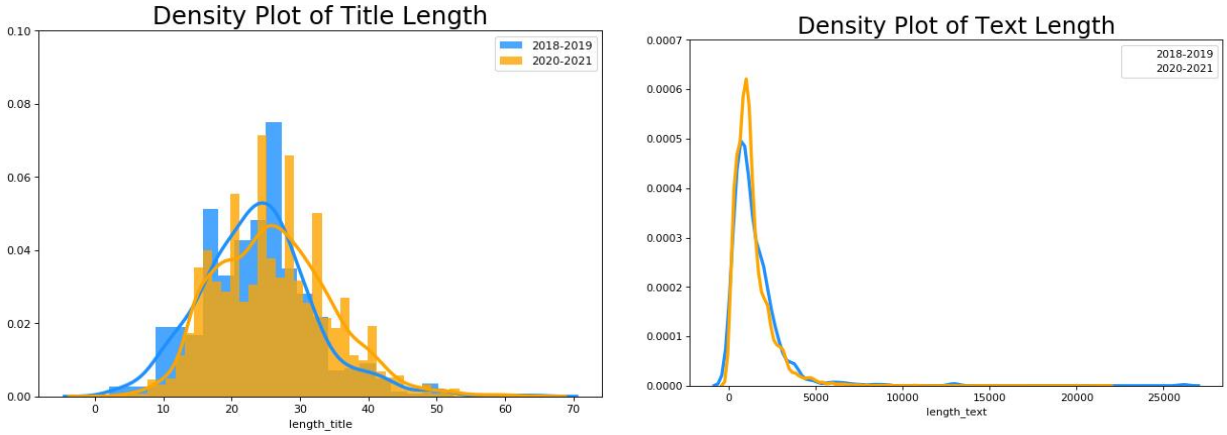


图 2. 文本标题和正文内容的长度分布

可以看到，文本标题的分布较为均匀，呈现出一定的类似正态分布特征，而文本内容长度则从 0-25000 变化不等，长度不一并且呈现严重的偏峰分布并且概率密度都很小，不利于建模。因此，我们认为，对标题赋予更高的权重会比纯粹用文本内容进行建模效果更优。

而就公众号内容而言，我们也绘制了如图 3 所示的词云图分析其高频主题：

图 3. 微信公众号词云图，左为 2018-2019，右为 2020-2021

分析其高频词变化，可以看到，2020-2021 年对政府媒体和防疫工作的热度明显比 2018-2019 更高。2018-2019 更多的是以景点、旅游、文化为主体的旅游，而 2021 年作为建党百年，有关当地政府政治经济类话题显然更多一些。在 2020-2021 年由于疫情原因，“确诊”“防疫”等高频词也频繁出现在文本当中，提倡在旅游的过程中响应防疫政策健康出行，也在百年奋斗的关键节点发布更多

体现新中国精神风貌的特色文旅。在这一过程中，“编辑部”“日报”等媒体高频词越来越多，这说明媒体在这一过程中起到了越来越重要的作用。

除此以外，我们还分析了三种评论和游记的高频词，并绘制词云图如下：

(1). 2018-2019 餐饮评论

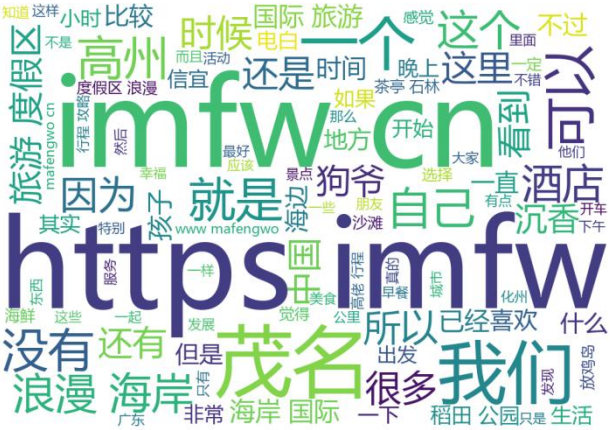
(2). 2020-2021 餐饮评论

(3). 2018-2019 酒店评论

(4). 2020-2021 酒店评论

(5). 2018-2019 景区评论

(6). 2020-2021 景区评论



(7). 2018-2019 游记攻略

(8). 2020-2021 游记攻略

图 4. 相关 UGC 数据和 OTA 数据的词云图

从餐饮评论的词云图中可以看出无论是 2018-2019 还是 2020-2021，游客对茂名旅游的餐饮评价总体都是不错的，给出了相当程度的好评。而在这些评论当中，菜品的味道和质量是最关键的一项评价因素。在 2018-2019 当中我们也发现如“环境”、“服务”也是评价的重要角度。就总体而言，“蛋糕”“榴莲千层”成为了词频最高的菜品，这和茂名的地理特征密不可分。

从酒店评论的词云图中可以看到酒店评论的整体情况是比较好的，大都给予了不错的评价。在酒店评论当中，“服务”则占了很高的一项比重，这也是现代旅游服务业发展所揭示的一点。而在服务领域，“前台”、“早餐”等也被频繁提及，构成对酒店评价的重要因素。

从景区评论的词云图中可以看到对景区而言“好玩”是最主要的，沙滩、公园、温泉等景点能够大大促进当地旅游业发展。除茂名以外，湛江也作为高频地名频繁出现，给游客带来不错的体验。

从游记攻略的词云图中可以看到，游记中综合出现了很多地名。包括高州、信宜等，而“浪漫”等气氛感则很能影响游客的游玩体验。网址的高频出现，说明网络在旅游信息传播当中有着重要的因素。

2.2 数据集探索

除了对数据集进行了如上的初步分析以外，我们还对数据进行了一些深入探究，将成果分享如下：

我们注意到游记攻略中存在出行天数和成本等属性，以出行天数和成本作为横纵坐标我们绘制了如图 5 所示的散点图：

图 5. 出行天数和人均费用散点图与直方图

从图 5 中我们可以看出出行天数与人均费用并没有很明显的相关关系，游客出行大多 1-2 天，平均开销不超过 1000 元大概 800 左右，是比较经济实惠的。一些游客也可能在短期内花费比较多，说明茂名旅游可以提供不同类型的服务，能够满足不同游客在不同价位服务上的需求。但出行日期和旅游人均费用都存在一些离群点，说明部分游客也有着个性化的需求。

与此同时，我们利用 SnowNLP 工具对不同评论进行情感分析并进行整合得到其情感得分的密度分布直方图如图 6 所示：

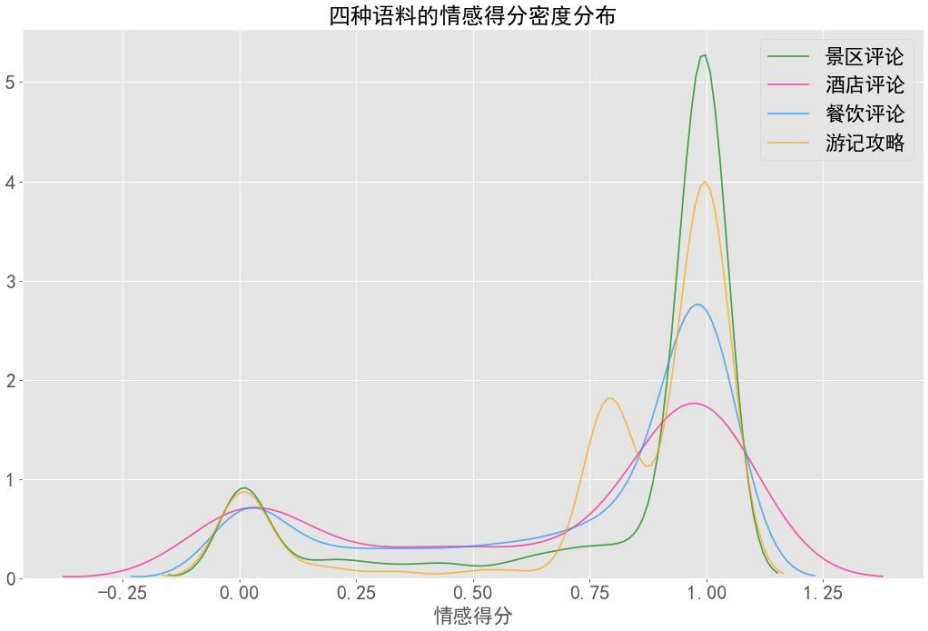


图 6. 情感得分的密度分布

SnowNLP 的给分是 0-1 之间的情感分数，若得分越低则表示情绪越负面，越高则评价越好。可以看到对茂名市旅游的评论语料中积极情绪占比居多，尤其是景区评论在高分区域非常集中。次之为游记攻略，并且游记攻略在 0.8 附近出现了一个新的峰值，说明总体而言游记攻略评价不仅好同时也比较客观，在好评的同时也给出了不足。餐饮评论分布更分散一些，而酒店评论在 0 附近分布最多，说明以酒店和民宿为主体的旅游住宿环境是茂名旅游的主要短板。

三. 模型假设与符号约定

3.1 模型假设

为了使得讨论结果更加合理且具有现实意义，我们将针对这一问题背景提出如下假设：

1. 由于原始数据是没有显式标注的，我们可以基于正则表达式和其他开源 NLP 工具对其进行初步标注，在经过检验以后得到更为精准的标注。

2. 为了对数据不同维度进行合理评分，我们认为，可以按照包含该维度关键词的子句情感评分代替这一维度的水平。例如，对环境的评价可以抽取包含“环境”、“气氛”等关键词的子句进行情感评分作为对环境这一水平的评价

3. 允许使用其他开源的评论评分数据集以及在这些数据集上训练的模型进行迁移学习，这也同样是由于缺乏维度标注。

3.2 符号假设

符号	说明
R	归一化的决策矩阵
r_{ij}	决策矩阵中第 i 个决策项第 j 个指标
w_j	熵权法求解的第 j 个指标的权重系数
R^+, R^-	正理想解和负理想解
D^+, D^-	决策项到正理想解和负理想解的距离
p_{ij}	对于指标 j，第 i 类所占比例
e_j	指标 j 的熵值
n	指标 j 有多少类
X,Y	两个不同的频繁项集
σ	LSTM 中表示 sigmoid 激活函数，Apriori 中表示元素数
f	神经网络的激活函数
W	神经网络的权重矩阵
b	神经网络的偏置项
o,h	LSTM 的隐藏状态
x,y	条件随机场中的两条状态序列

（注：若有变量未列入表内或与表内释义不一致请以原文解释为主）

四. 研究方法综述

4.1 基于深度学习的文本分类算法

文本分类是一类典型的监督学习任务，它基于文本内容和文本的标签训练分类器并在测试数据上进行验证。文本分类的基本思想是先将文本利用一些嵌入手段（例如 TF-IDF[1]、N-Grams[2]、Word2Vec[3]、GloVe[4]等）进行向量化，然后转化为机器学习中的分类任务用分类器求解。传统的文本分类算法包括 SVM[5]、朴素贝叶斯[6]等，但随着深度学习的发展，深度学习方法在文本分类领域的应用已经被证明比传统机器学习方法更有效[7]。这一部分我们将对比几种常见的文本分类模型。

4.1.1 不引入注意力机制的文本分类

卷积神经网络的结构已经在图像处理上得到了广泛应用，能够高效提取图像的特征。将卷积神经网络架构引入文本分类任务时一个常用的模型就是 TextCNN，TextCNN 网络是 2014 年提出的用来做文本分类的卷积神经网络，由于其结构简单、效果好，在文本分类、推荐等 NLP 领域应用广泛[8,9]。TextCNN 的结构比较简单，输入数据首先通过一个 embedding layer，得到输入语句的 embedding 表示，然后通过一个 convolution layer，提取语句的特征，最后通过一个 fully connected layer 得到最终的输出。图 7 为 TextCNN 的模型图

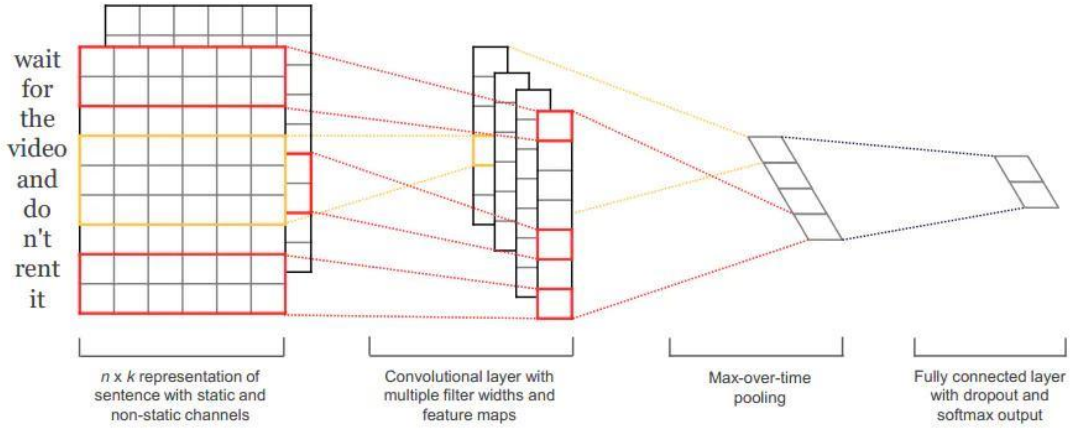


Figure 1: Model architecture with two channels for an example sentence.

图 7. TextCNN 的模型架构

TextCNN 最大优势网络结构简单 ,在模型网络结构如此简单的情况下，通过引入已经训练好的词向量依旧有很不错的效果。其网络结构简单导致参数数目少，计算量少，训练速度快，TextCNN 是很适合中短文本场景的强 baseline，但不太适合长文本，因为卷积核尺寸通常不会设很大，无法捕获长距离特征。同时 max-pooling 也存在局限，会丢掉一些有用特征。

循环神经网络也常被用于文本等序列模型的建模中。文本分类问题就是对输

入的文本字符串进行分析判断，之后再输出结果，但字符串无法直接输入到 RNN 网络，因此在输入之前需要先对文本拆分成单个词组，将词组编码成一个向量，每轮输入一个词组，得到输出结果也是一个向量[10]。嵌入表示将一个词对应为一个向量，再最后进行全连接操作对应到不同的分类即可。RNN 进行文本分类时将问题抽象为序列，最后使用 softmax 进行分类预判。但 RNN 网络不可避免地带来问题就是最后的输出结果受最近的输入较大，而之前较远的输入可能无法影响结果，也就是造成“信息瓶颈”。

RCNN 模型综合了 CNN 和 RNN 二者的特性，它使用双向循环网络捕捉上下文信息，又用最大池化等卷积模型特性捕捉全文关键信息，卷积层的特征提取的功能被双向 RNN 替代，因此整体结构变为了双向 RNN+池化层[11]。模型在经过嵌入过程得到词向量的嵌入表示以后利用前向-后向循环网络得到上下文的表示，再将词向量与上下文表示联合为一个新的向量。向量经过全连接层以后再进行最大池化即可将变长的文本转换为固定长度的向量。随后将这个句子的表示向量喂进一个全连接 softmax 层进行分类概率预测。这一模型在高准确率的同时速度也较快，但在大规模文本上注意力机制的效果会比传统 RNN、CNN 类模型更好。

2017 年腾讯提出了 DPCNN 模型，DPCNN 和 ResNet[16]的结构是非常相似的，通过引入 shortcut 机制有效解决了梯度消失问题，通过不断加深网络，可以抽取长距离的文本依赖关系。实验证明在不增加太多计算成本的情况下，增加网络深度就可以获得最佳的准确率[15]。DPCNN 在文本分类领域取得比传统 TextCNN 更好的效果，但由于其深度过高训练难度相较 TextCNN 也更大。

4.1.2 引入注意力机制的文本分类模型

2017 年谷歌团队提出的 Transformer 模型打破了传统卷积或序列对齐的循环结构进行特征表示与抽取的模式，创新性地引入注意力机制解决问题，并且其自带的多头注意力能够对不同形式的数据进行特征融合与表示，适合多模态问题处理[12]。

使用深度学习进行文本分类时，最终池化时，max-pooling 通常表现更好，但到更细粒度的分析时，max-pooling 可能又把有用的特征去掉了，这时便可以用 attention 进行句子表示的融合[13]。一个经典的文本分类模型就是 HAN(Hierarchy Attention Network)，它首先将文本分为句子、词语级别，先对每个句子用 BiRNN+Att 编码得到句向量，再对句向量用 BiRNN+Att 得到文档级别的表示进行分类[14]。HAN 模型针对篇章文本由单词组成句子，再由句子组成文章的特点，从句子层面到篇章层面分别建模。模型结构科学合理，并且增加了 Attention 机制，性能相对 transformer 有更大改进。

随着深度学习的技术发展，动态的词嵌入方法，尤其是基于大规模预训练模型的文本向量化方法发展如火如荼。2018 年底微软提出的 BERT(Bidirectional Encoder Representation from Transformers)相较于 Elmo[17]和 GPT-2[18]取得了更好的表现，目前也是应用最广泛的文本向量化方法之一,因为在 BERT 中，特征提取器也是使用的 Transformer,且 BERT 模型是真正在双向上深度融合特征的语言模型[19]。

如图 7，不同于 GPT、ELMo 模型，BERT 采用的是 Transformer Encoder，也就是说每个时刻的 Attention 计算都能够得到全部时刻的输入。BERT 预处理进行

下游任务的输入是三个嵌入表示叠加，如图 8，Token embedding 表示当前词的 embedding，Segment Embedding 表示当前词所在句子的 index embedding，Position Embedding 表示当前词所在位置的 index embedding。

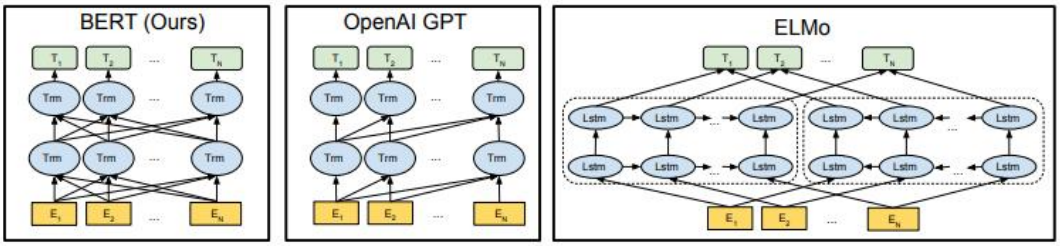


图 8. 几种大规模预训练模型的对比

在 BERT 出现之前的词嵌入技术中，一个句子的嵌入表示，往往简单地使用各个单词的词嵌入表示进行平均或加和得到，这就导致无法得到包含深层语义和语序信息的词嵌入表示，实际任务中效果也较差。而通过 BERT 得到的词嵌入表示融入了更多的语法、词法以及语义信息，而且动态的改变词嵌入也能够让单词在不同语境下具有不同的词嵌入表示。如图 9 所示为 BERT 的嵌入过程。

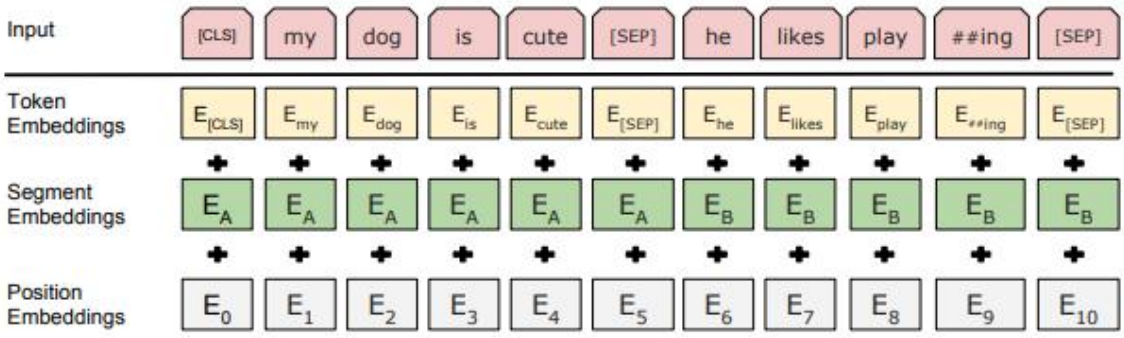


图 9. BERT 嵌入方法

BERT 模型虽然考虑了上下文语义，但是缺少了知识信息。为此，百度提出 ERNIE 模型，在 BERT 的基础上引入知识驱动，利用了知识图谱中的多信息实体（informative entity）来作为外部知识改善语言表征[20]，模型采用 TransE[21]知识嵌入算法，将编码后的知识信息整合到语义信息当中；为了将更好地将语义和知识信息融合起来，模型改进了 BERT 模型的架构，并设计了新的预训练任务，这样就可以将实现知识信息与语义信息的融合。ERNIE 模型在中文文本分类上表现比 BERT 更优，并且可以做到一定程度的文本理解，效果相比 BERT 可能更好。

4.2 命名实体识别方法

命名实体是指可以认为是某一个概念的实例。命名实体识别的目的是一种序列标注问题,(Named EntitiesRecognition，NER)，就是识别这些实体指称的边界和类别，主要关注人名、地名和组织机构名这三类专有名词的识别方法[22]。传统的命名实体识别方法依赖于手工规则的系统，结合命名实体库，对每条规则进行权重辅助，然后通过实体与规则的相符情况来进行类型判断。大多数时候，规则往往依赖具体语言领域和文本风格，难以覆盖所有的语言现象。典型应用形如在航空领域的命名实体识别[23]，军事领域的命名实体识别等[24]，对特定实体建立词库进行检索判断，尽管在各自特定领域准确率较高但效率较低并且难以形成

通用实体识别方法。

4.2.1 BiLSTM-CRF 模型

应用于 NER 中的 BiLSTM-CRF 模型主要由 Embedding 层，双向 LSTM 层，以及最后的 CRF 层构成，实验结果表明 BiLSTM-CRF 已经达到或者超过了基于丰富特征的 CRF 模型，成为目前基于深度学习的 NER 方法中的最主流模型[32,33]。在特征方面，该模型继承了深度学习方法的优势，无需特征工程，使用词向量以及字符向量就可以达到很好的效果，如果有高质量的词典特征，能够进一步获得提高。图 10 为 BiLSTM-CRF 的示意图。

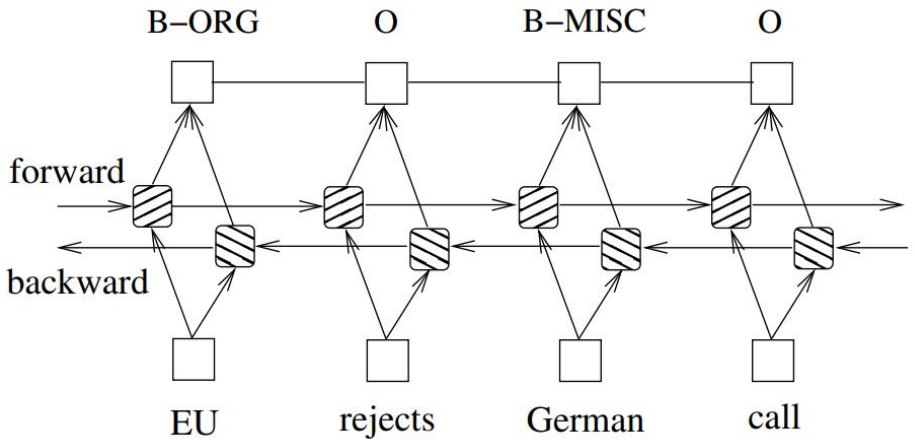


图 10. BiLSTM-CRF 的结构示意

首先，将句子 x 中的每个单词表示为一个向量，其中包括单词的嵌入和字符的嵌入。字符嵌入式随机初始化的。词嵌入通常是一个预训练的词嵌入文件导入的。所有的嵌入将在训练过程中进行微调。其次，BiLSTM-CRF 模型的输入是这些嵌入，输出是句子 x 中的单词的预测标签。我们将模型分为两层介绍：

BiLSTM 层

BiLSTM 是将正反两个方向的 LSTM 拼接在一起进行序列预测，由于自然语言处理中需要综合考虑上下文信息所以需要引入双向 LSTM 对上下文语义同时进行建模[29]。

LSTM 结构如图 11 所示[30]：

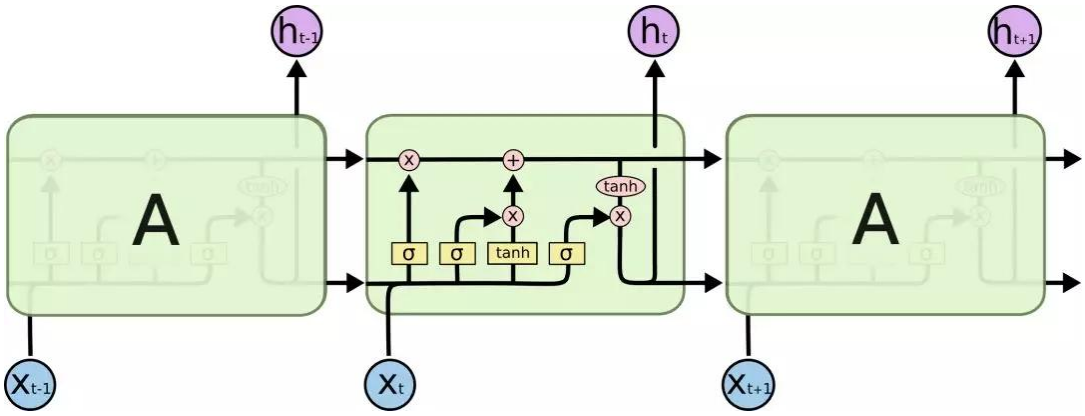


图 11. LSTM 结构示意图

图 11 最上方的 c_{t-1} 到 c_t 的总线贯穿始终，是整个 LSTM 网络的核心。在总线的下方是三个门。从左到右的三个门分别为遗忘门，输入门和输出门。

对于遗忘门，它的作用是将接收过的信息进行选择性地遗忘，可以主动调节不同位置信息的作用大小。对此，我们有：

$$f_t = \sigma(W_f(h_{t-1}, x_t) + b_f) \tag{1}$$

而输入门的作用是更新单元的状态。将新的信息有选择性地输入来代替被遗忘的信息，并生成候选向量 C。下面的方程解释了输入门生成的候选向量：

$$i_t = \sigma(W_i(h_{t-1}, x_t) + b_i) \tag{2}$$

$$C = \tanh(W_C(h_{t-1}, x_t) + b_C) \tag{3}$$

输出门可以给出结果，同时将先前的信息保存到隐层中去。同样的，我们有：

$$o_i = \sigma(W_o(h_{t-1}, x_t) + b_o) \tag{4}$$

$$c_t = f_t c_{t-1} + i_t C \tag{5}$$

$$h_t = o_t \tanh(c_t) \tag{6}$$

我们可以看到，LSTM 宏观上也是关于 x_{t-1} 和 x_t 的函数，但是由于多了门控单元对长期信息和短期信息的不同处理模式，网络能够对先前的长期信息保持一定记忆，这克服了传统 RNN 只能针对先前的短期数据进行计算的缺点。

LSTM 网络比较适合于时间序列的中短期预测，对于长期预测效果仍然欠佳。问题一的短期时间预测可以利用 LSTM 进行调参与测试。而 BiLSTM 有一条正向传播线一条反向传播线，当前步的过去和未来输入都可以作用于损失函数[31]。训练的话，就先分别训练两条线，当成两个 RNN 训练，再组合。此外，BILSTM 的当前输出与前向链输出和后向链输出有关，他们俩受到外界激励的作用。

条件随机场层

McCallum 等 2003 年最先将条件随机场（CRF）模型用于命名实体识别。由于该方法简便易行，而且可以获得较好的性能，因此受到业界青睐，已被广泛地应用于人名、地名和组织机构等各种类型命名实体的识别，并在具体应用中不断得到改进，可以说是命名实体识别中最成功的方法[26]。

基于 CRF 的命名实体识别将其看作一个序列标注问题。其基本思路是将给定的文本首先进行分词处理，然后对人名、简单地名和简单的组织机构名进行识别，最后识别复合地名和复合组织机构名。CRF 是给定一组输入序列的条件下，另一组输出序列的条件概率分布模型[27]。一组随机变量按照某种概率分布随机赋值到某个空间的一组位置上时，这些赋予了随机变量的位置就是一个随机场[28]。而条件随机场，就是给定了一组观测状态下的马尔可夫随机场，这个随机场满足马尔可夫性。也就是说 CRF 考虑到了观测状态这个先验条件，这也是条件随机场中的条件一词的含义。条件随机场的数学模型为：

$$P(Y_v | X, Y_w, w \neq v) = P(Y_v | X, Y_w, w \sim v) \tag{7}$$

这一表达式充分体现了条件随机场所带有的马尔可夫性。

CRF 中有两类特征函数，分别是状态特征和转移特征，状态特征用当前节点的状态分数表示，转移特征用上一个节点到当前节点的转移分数表示。CRF 损失

函数的计算，需要用到真实路径分数和其他所有可能的路径的分数。这里真实路径表示真实的词性序列，其他可能的路径表示其他的词性序列。在给定某个状态序列时，某个特定的标记序列概率为：

$$P(Y|X)=\frac{1}{Z}\exp\left(\sum_{j}\sum_{i=1}^{n-1}\lambda_j t_j(y_{i+1},y_i,x,i)+\sum_{k}\sum_{i=1}^{n}\mu_k s_k(y_i,x,i)\right) \tag{8}$$

其中，t 和 s 分别为转移特征函数和状态特征函数，Z 为规范化因子。

条件随机场为命名实体识别提供了一个特征灵活 、全局最优的标注框架 ，但同时存在收敛速度慢 、训练时间长的问题 。相比而言，同为机器学习方法的隐马尔可夫模型则更加迅速，处理效率更高，往往用在精度要求不是太高但要求实时标记的场合。

CRF 层可以向最终的预测标签添加一些约束，以确保它们是有效的。这些约束可以由 CRF 层再训练过程中从训练数据集自动学习。有了这些有用的约束，无效预测标签序列的数量将显著减少。

BERT-BiLSTM-CRF 模型在输入 BiLSTM 模型之前先用 BERT 模型对词向量进行了嵌入与建模，由于 BERT 模型的高效性使得后续过程得到显著提升。

目前 BERT-BiLSTM-CRF 模型已经证实在多种中文语料的命名实体识别任务中得到显著提升。包括中药文本命名实体识别[34]、人民日报语料库、MSRA 语料[35]等，性能上相比 BiLSTM-CRF 有更显著的提升。

4.3 关联关系挖掘方法

关联规则挖掘本质上是为了发现频繁项集及其之间的关联规则，在这一问题中我们常用“购物篮模型”对问题进行建模。尽管关联规则是一种更为复杂的建模，但本质上还是基于频繁项集的发现。

4.3.1 Apriori 算法

Apriori 算法是最经典的关联关系挖掘算法，通过逐层迭代剪枝的方法分别采用支持度和置信度来量化频繁项集和关联规则[36]。

在一个购物篮模型中，我们用项描述所有种类的产品，每一个产品可以称为一个项，而一个或一组产品在一个语料中同时出现我们将其构成一个集合，称为项集。若给定两个项集 XY，定义其支持度和置信度分别为：

$$\begin{aligned} support(X\rightarrow Y)&=\frac{\sigma(X\cup Y)}{N}\\ confidence(X\rightarrow Y)&=\frac{\sigma(X\cup Y)}{\sigma(X)}\end{aligned} \tag{9}$$

给定事务的集合 T，关联规则发现是指找出支持度大于等于 minsup (最小支持度)并且置信度大于等于 minconf(最小置信度)的所有规则，minsup 和 minconf 是对应的支持度和置信度阈值。关联规则的挖掘是一个两步的过程：

(1)频繁项集产生：其目标是发现满足最小支持度阈值的所有项集（至少和预定义的最小支持计数一样），这些项集称作频繁项集。

(2)规则的产生：其目标是从上一步发现的频繁项集中提取所有高置信度的

规则，这些规则称作强规则。（必须满足最小支持度和最小置信度）

Apriori 算法会自底而上搜索，先按照预先设置的最小支持度筛选出频繁 1 项集，再根据频繁 1 项集生成频繁 2 项集，一直到频繁 k 项集位置。至此发现频繁项集。而关联规则的挖掘则根据产生的频繁项集计算每个频繁项的子集与它补集之间的置信度，若满足最小置信度要求则输出关联规则。

算法 1. Apriori 算法

```
Algorithm Apriori(T)
1   C1 ← init-pass(T);                          // the first pass over T
2   F1 ← {f | f ∈ C1, f.count/n ≥ minsup};      // n is the no. of transactions in T
3   for (k = 2; Fk-1 ≠ ∅; k++) do               // subsequent passes over T
4     Ck ← candidate-gen(Fk-1);
5     for each transaction t ∈ T do             // scan the data once
6       for each candidate c ∈ Ck do
7         if c is contained in t then
8           c.count++;
9       endfor
10    endfor
11    Fk ← {c ∈ Ck | c.count/n ≥ minsup}
12  endfor
13  return F ← ⋃k Fk;
```

4.3.2 改进的 PCY 算法

Apriori 算法的一个重要缺陷就是花费时间太多。而中国科学家 Park, Chen, Yu 设计的 PCY 算法则可在大数据情境下实现高效快速的关联关系挖掘[37]。其主要创新点有：

1. 将哈希函数应用到频繁项挖掘中；

2. 第一次扫描事务数据库时，将剩余的空间存放哈希表，从而降低第二次扫描时占用的大量空间。

PCY 算法首先第一次扫描事务数据库，并进行计数，根据阈值筛选出频繁 1-项集；然后在第一次扫描的同时，对事务集中每个事务中各个项进行两两组合，并通过哈希函数映射到相应的桶中，对相应的桶计数进行更新；在下一次扫描之前，首先根据每个桶中的计数来判断是否为频繁桶，频繁桶是指计数不低于某个阈值的桶。频繁的桶对应的位图记做 1，否则记做 0；第二次扫描事务数据库时根据频繁 1 项集生成一系列的 2 项组合，分别通过哈希函数获得其桶的编号，并根据位图来判断。如果对应的位图值为 1，则将其保留，否则剔除。因此最后保留的就是候选 2 项集，对其进行计数并筛选频繁 2 项集。更高阶频繁项集的生成方法类似。

4.4 知识图谱构建技术

4.4.1 知识图谱的概念、发展及应用

在大数据时代，知识工程是从大数据中自动或半自动获取知识，建立基于知识的系统，以提供互联网智能知识服务[38]。而知识图谱（Knowledge Graph）以

结构化的形式描述客观世界中概念、实体及其之间的关系，提供了一种更好地组织、管理和理解互联网海量信息的能力[39]。知识图谱在知识融合、语义搜索和推荐、问答和对话系统、大数据分析与决策等领域中已经凸显出越来越重要的应用价值，将知识组织成图的形式，能够进行认知推理也更容易可视化。

4.4.2 基于图数据库 Neo4j 的知识图谱构建与可视化

Neo4j 是一个高性能的 NOSQL 图形数据库，它将结构化数据存储在网络上而不是表中[40]。它将结构化数据存储在拓扑图上而不是表中，同时也可以被看作是一个高性能的图引擎，该引擎具有成熟数据库的所有特性。Neo4j 因其嵌入式、高性能、轻量级等优势，越来越受到关注，将数据存储在 Neo4j 中安全可靠。

4.5 TOPSIS 评价模型

评价类问题常用的方法有层次分析法（AHP）、模糊综合判别法、灰色关联分析（GRA）、主成分分析（PCA）和优劣解距离法（TOPSIS）等[41]，但这些方法侧重点各有不同，前两者属于主观赋权评价法，在该问题中由于缺乏运输订购类问题中相关变量的权重分配研究，进行应用时权重会带有一定主观性和随机性，不利于量化。灰色关联分析更多用于分析序列之间的相关性而非重要性，主成分分析更多则用于降维减少变量维度，并且要求构建的主成分有充分的可解释性，在这一问题中不能很好地达到我们的目的，而 TOPSIS 方法是多目标决策分析中一种常用的有效方法，是一种趋近于理想解的排序法。它根据有限个评价对象与理想化目标的接近程度进行排序，在现有的对象中进行相对优劣的评价[42]，故我们选用 TOPSIS 方法进行求解。

TOPSIS 评价法是有限方案多目标决策分析中常用的一种科学方法，其基本思想为，对原始决策方案进行归一化，然后找出最优方案和最劣方案，对每一个决策计算其到最优方案和最劣方案的欧几里得距离，然后再计算相似度。若方案与最优方案相似度越高则越优先。基本流程如下

Step1：根据归一化得到的决策矩阵（这里我们选取 min-max 归一化）和权重向量构造规范化权重矩阵 R

$$R=[w_j r_{ij}] \tag{10}$$

Step2：确定正理想解 R+和负理想解 R-,其中分别表示效益型指标和成本型指标：

$$\begin{aligned} R^{+}&=\left\{\max_{j\in J^{+}} r_{ij},\min_{j\in J^{-}} r_{ij}\right\} \\ R^{-}&=\left\{\max_{j\in J^{+}} r_{ij},\min_{j\in J^{-}} r_{ij}\right\} \end{aligned} \tag{11}$$

Step3：计算各评价对象 i（i=1,2,3…，402）到正理想解和负理想解的欧几里得距离 Di：

$$D_i^+ = \sqrt{\sum_{j=1}^{n} [r_{ij} - R_j^+]^2}$$
$$D_i^- = \sqrt{\sum_{j=1}^{n} [r_{ij} - R_j^-]^2} \tag{12}$$

Step4：计算各评价对象的相似度 Ci

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{13}$$

可以看到，相似度是与负理想解与两理想解距离之和的比值，若占比越大则说明离负理想解越远，越优先选择。

Step5：根据 Ci 大小排序可得到结果。

权重向量的构建是 TOPSIS 应用的核心，需要尽可能削弱其主观性。这里我们使用熵权法构建权重向量。熵权法基于信息论，基于信息论的熵值法是根据各指标所含信息有序程度的差异性来确定指标权重的客观赋权方法，仅依赖于数据本身的离散程度[43]。熵用于度量不确定性，指标的离散程度越大，说明不确定性越大，则最终熵值越大，该指标值提供的信息量越多，则权重也相应越大。

根据信息论中对熵的定义[44]，熵值 e 的计算如下所示

$$e_j = -\frac{\sum_{i=1}^{n} p_{ij} \ln p_{ij}}{\ln n} \tag{14}$$

式子(4)中代表对于某一个属性 j，第 i 类占样本的比例。n 为属性 j 的取值数量。所以权重系数 w 定义为：

$$w_j = \frac{1 - e_j}{\sum_{i=1}^{m} (1 - e_i)} \tag{15}$$

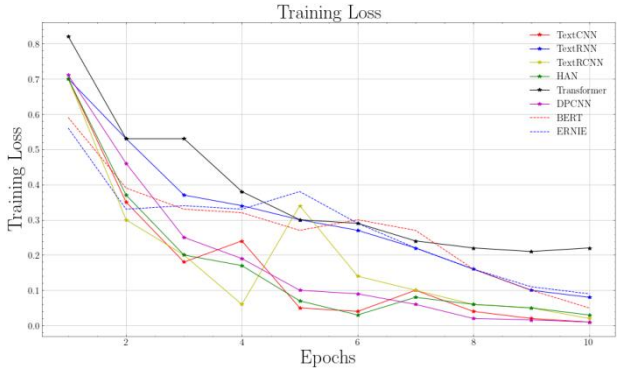
这样，我们构建了 TOPSIS 综合评价与熵权分析法的综合模型。

五. 模型建立与求解

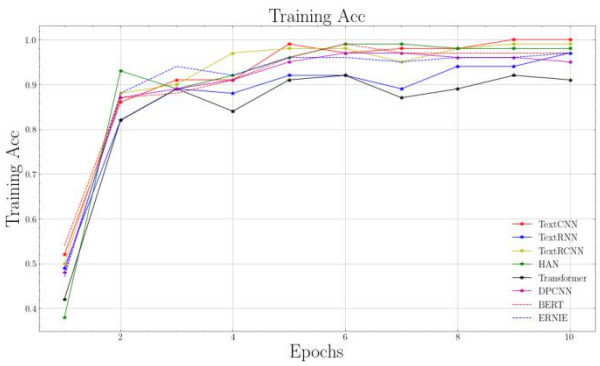
5.1 问题一的模型

问题一是一个典型的文本分类问题。这里我们对比了八种不同的模型，我们以将文本标题分类赋予 0.8 的权重而将文本内容赋予 0.2 的权重进行模型综合。文本的标签基于问题提供的关键词通过正则表达式检索的方式获得。我们按照 8:1:1 的比例切分训练集、验证集和测试集，对不同的模型进行了训练。

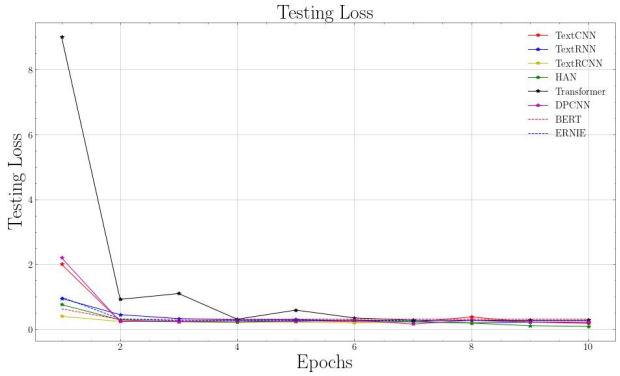
我们可以对比不同网络在训练过程中的误差损失和训练准确率变化，将训练和验证过程的误差损失和准确率随迭代轮数的变化曲线绘制在图 12 中：



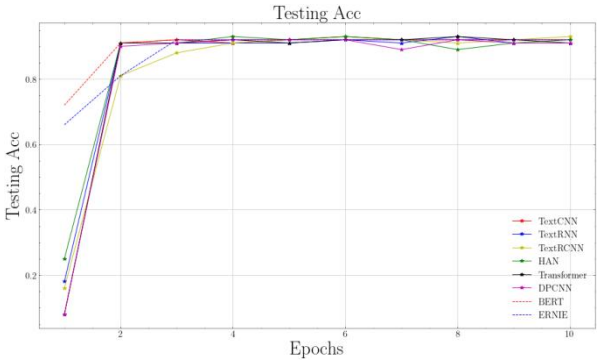
(1). 训练损失



(2). 训练准确率



(3). 验证损失



(4). 验证准确率

图 12. 损失函数与准确率随训练轮数的变化曲线

从图 12 中我们可以看到，几种网络整体的训练损失和验证损失在训练 10 轮以后基本都已经收敛，训练准确率和验证准确率整体上也都比较高。但 TextCNN 的损失最低，准确度也是八种不同模型当中最高的。相反，在多项 NLP 任务中取得 SOTA 的 Transformer 系列模型以及大规模预训练模型在这一问题中表现非常不鲁棒，取得的准确率也非常低。

为了对不同文本分类模型进行合理的效果评估，我们将八个模型的 AUC 曲线绘制在图 13 中。可以看到，TextCNN 的 AUC 值是最高的，而在文本领域取得大规模应用的 BERT、ERNIE 等模型表现反而比较差。这说明对于这一类文本分类问题而言，复杂的模型往往不是最好的。在某些场合下使用较为简单的模型反而会有更好的效果。

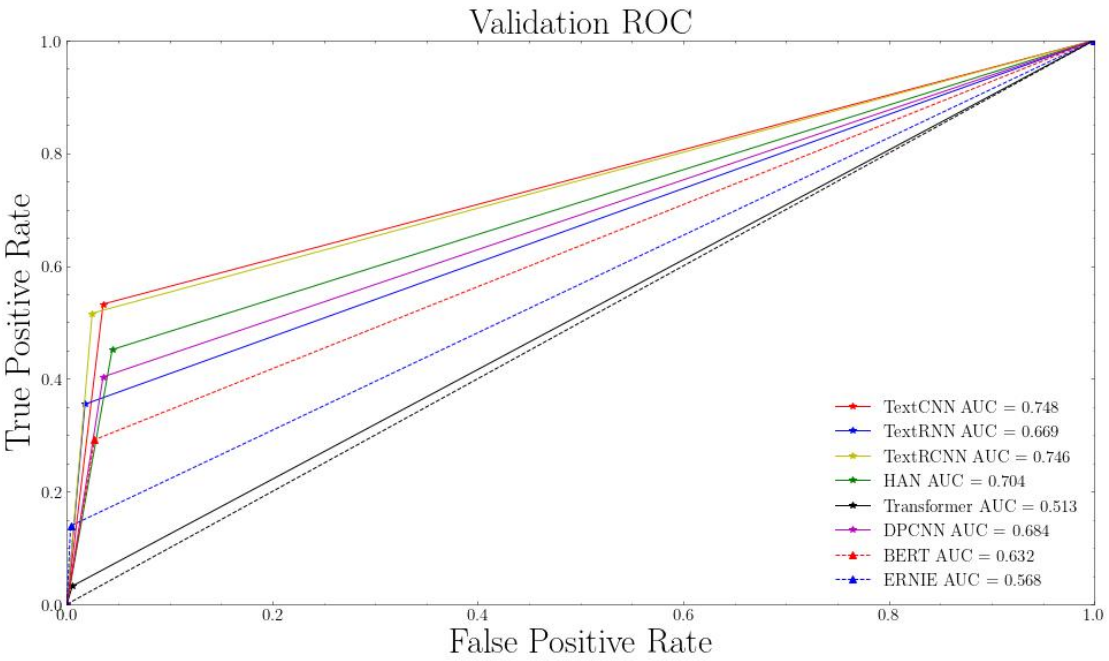


图 13. 分类的 AUC 曲线

不同模型的表现如表 1 和图 14 所示。

表 1. 不同模型的性能对比

model	time	accuracy	f1	auc
TextCNN	25	0.9211	0.7655	0.748
TextRNN	22	0.9205	0.7126	0.669
TextRCNN	12	0.93	0.7772	0.746
HAN	21	0.93	0.7177	0.704
DPCNN	17	0.9094	0.7089	0.684
Transformer	20	0.8998	0.5034	0.513
BERT	132	0.8855	0.6527	0.632
ERNIE	140	0.8998	0.6269	0.568

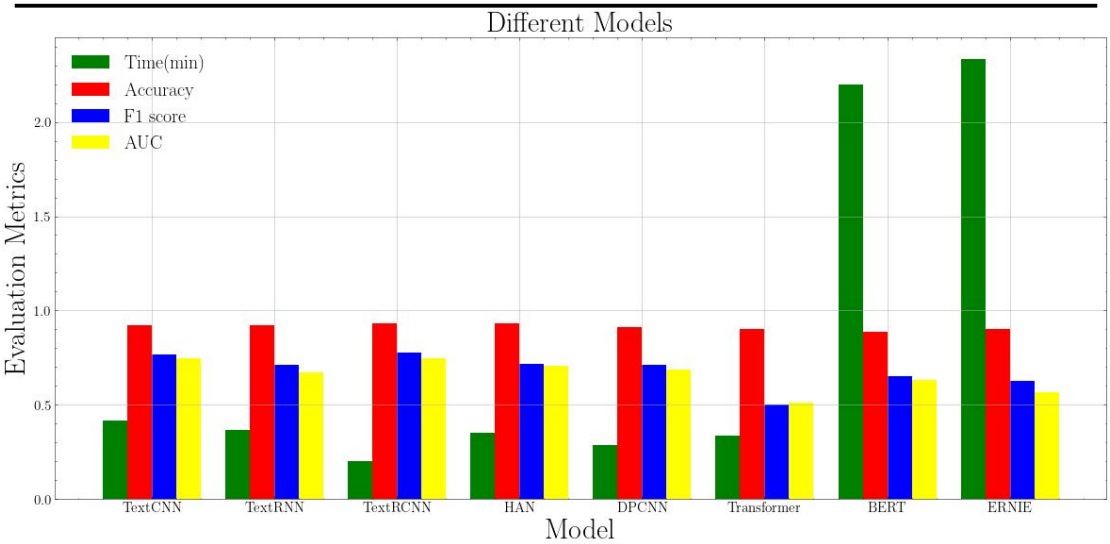


图 14. 衡量指标的可视化

从图中可以看出，从训练时间而言，使用大规模预训练模型的用时都已经超

过了 1 分钟，速度较慢的同时效果也较差。从准确率、F1 分数和 AUC 值来讲，TextCNN 无疑是表现最优秀的，时间差距也与其他模型相比毫不逊色，所以我们选择使用 TextCNN 进行文本分类。最终结果写入 result1.csv 中。

5.2 问题二的模型

5.2.1 基于命名实体识别的产品抽取

由于原始数据集并没有给出产品的标注，我们基于 TexSmart 对文本进行了初步命名实体识别标注后进行再对比几种不同的模型进行训练。TexSmart 是腾讯开发的大规模自然语言处理平台，可以进行文本分类、分词、词性标注、命名实体识别、语法分析、词汇知识图谱等内容[45]，能够初步辅助进行自然语言理解工作。

在经过 TexSmart 标注命名实体以后我们按照实体对应的分类进行产品归类得到了目标产品的初步标注。基于多种模型我们的命名实体识别效果如表 2 所示：

表 2. 命名实体识别效果

Evaluation	HMM	CRF	BiLSTM	BiLSTM+CRF	BERT-BiLSTM+CRF
Recall	91.22%	95.43%	95.32%	95.72%	95.65%
Precision	91.49%	95.43%	95.37%	95.74%	95.69%
F1-score	91.30%	95.42%	95.32%	95.70%	95.64%

可以看到就总体效果而言 BiLSTM+CRF 相比 BERT+BiLSTM+CRF 更高一些，这也和任务一的一项发现类似，在小批量短文本的情况下 BERT 的效果并不一定会比传统的序列对齐 RNN 或卷积结构更好。因此我们选择使用 BiLSTM+CRF 进行序列和实体的标注。

对于提取出的实体分布，我们也绘制了如图 15 所示的条形图

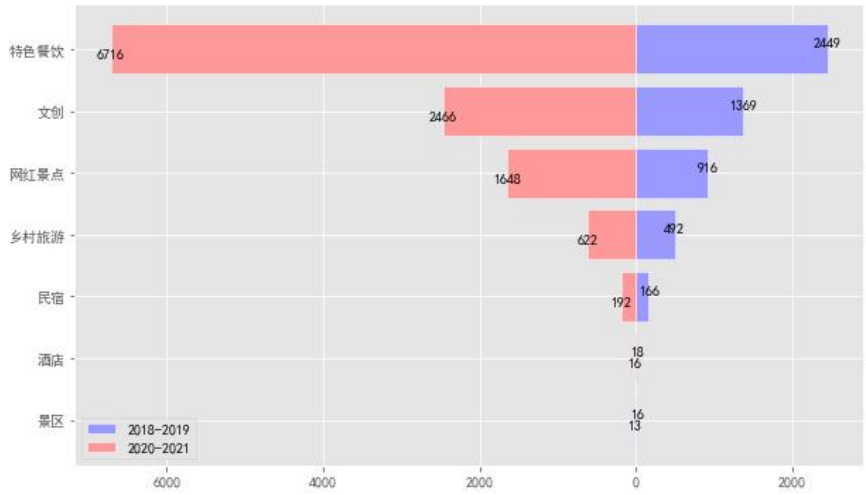


图 15. 产品实体在 2018-2019 和 2020-2021 的分布图

从图中可以看到特色餐饮产品是旅游中出现最多的，2018-2019 年出现 2449 项而 2020-2021 年出现 6716 项。其次是文创产品，2018-2019 年出现 1369 项而

2020-2021 年出现 2466 项，排名第二。网红景点则是第三位，2018-2019 为 916 项，2020-2021 为 1648 项。这三项构成文化旅游的主体也是很符合常理的。

5.2.2 基于 TF-IDF 与评分模型的热度评价

我们在开源的美团评论数据集[46]上训练了评论的打分模型，以 1-5 的 整数作为评分，利用 TextCNN 训练网络效果如图 16 所示：

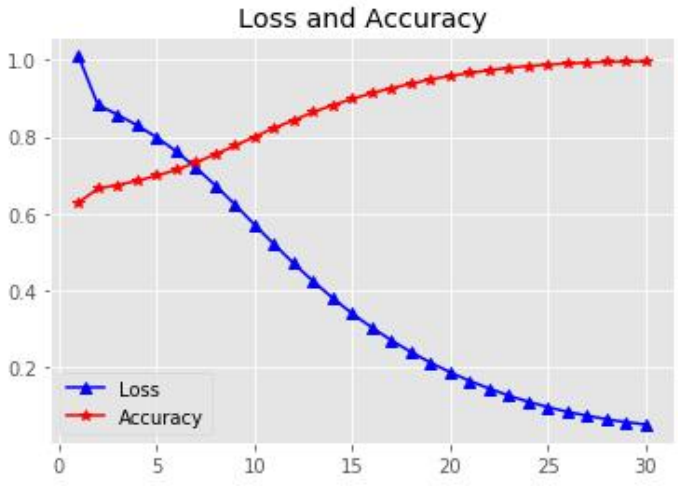


图 16. 评分神经网络的训练效果

从图中可以看到，评分的分类网络损失函数收敛且准确率较高，可以用于评分任务。我们基于这一模型对所有的 OTA、UGC 数据进行了评分。

与此同时，我们也对语料中出现的每一个产品计算了其 TFIDF 得分。我们将一个语料当中出现的所有产品实体看成一个列表，对于不同的产品可以通过词频统计的方式获得 TF，而通过语料又可以获得逆文档指数 IDF，两项乘积则可以描述产品在这一语料当中的重要性程度。

计算出评分和 TFIDF 值以后我们对其进行乘积，并进行了 min-max 标准化规约将其规约到 0-1 之间，规约形式如下：

$$x = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{16}$$

最终计算的热度分布如表 3 所示。

5.2.3 基于 TOPSIS 模型的综合热度评价

我们通过正则表达式，从每个语料当中选择了包含目标产品、包含环境、包含服务和包含成本的语句，利用 SnowNLP 工具对几种因素进行评分作为四项维度的打分以后便于进行指标分析。

利用 4.5 中建立的 TOPSIS 模型得到的热度如表 3 所示。

表 3. 两种热度评价下部分产品热度表格

产品名称	环境	服务	质量	成本	TOPSIS 得分	TFIDF 得分
花海	0.5	0.5	0.734869763	0.470042859	0.59811203	0.140413271

桂花	0.5	0.5	0.130567942	0.470042859	0.345577894	1
杨桃	0.5	0.5	0.130567942	0.470042859	0.345577894	0.213041424
石榴	0.5	0.5	0.144033573	0.470042859	0.349520714	0.505877841
西江温泉	0.5	0.5	0.064879657	0.5	0.333791741	0.689981975
玻璃栈道	0.5	0.5	0.180619718	0.5	0.366198258	0.512243391
天马山	0.5	0.5	0.081030306	0.256973202	0.299935822	1
石根山	0.5	0.5	0.966182832	0.5	0.672861935	0.02319727
树	0.5	0.5	0.875348055	0.5	0.650935076	0
地质公园	0.97	0.5	0.951234959	0.5	0.721339535	1
栈桥	0.5	0.5	0.998133558	0.5	0.678786224	0.088531132
花海	0.5	0.5	0.734869763	0.470042859	0.59811203	0.140413271
桂花	0.5	0.5	0.130567942	0.470042859	0.678786224	0.707106781

可以发现 TOPSIS 得分本身就处在 0-1 之间，做为热度更为合理。但利用 TFIDF 与评分模型的方法本身与几种属性得分无关，并且显得区分度更为明显，是两种不同的评价体系。两种评价模型各有优劣，但我们经过对比后认为 TOPSIS 得分更为合理。

5.2.4 情绪因素对热度的调节效应实证研究

在 5.2.2 的基础上，为了对结合 TOPSIS 和 TFIDF-评分模型的热度结果进行多维度评价与探究，我们利用 SPSSPRO 进行了情绪因素对热度的多维度调节效应实证研究。在这一过程中，我们认为用户的情绪会在四个维度对热度的模型当中产生影响，于是我们以情绪作为调节变量分别对四个维度进行了调节效应的研究。调节效应分析是为了探索情绪在这一过程中调节的效果好坏，在何时起到对热度评分的调节作用。首先，我们分别按照环境、服务、产品、成本这四项因素评分与热度之间进行 OLS 一元回归，再进行了四元最小二乘回归，回归方程形如：

$$\begin{cases} y = 0.239 + 0.047\,\text{quality} \\ y = 0.249 + 0.023\,\text{cost} \\ y = 0.053 + 0.359\,\text{environment} \\ y = 0.142 + 0.212\,\text{service} \\ y = -0.049 + 0.222\,\text{service} + 0.367\,\text{environment} + (-0.039)\,\text{quality} + (-0.013)\,\text{cost} \end{cases} \tag{17}$$

我们对环境、服务、产品质量和成本四个维度与热度评分进行 OLS 最小二乘回归和情绪因素对其调节效应检验。表 4-表 7 分别为四个因素的假设检验表格：

表 4. 情绪对环境与热度的调节效应

	模型1				模型2				模型3			
	B	标准误	t	p	B	标准误	t	p	B	标准误	t	p
const	0.053	0.031	1.741	0.083	0.03	0.062	0.485	0.628	0.292	0.149	1.963	0.05
environment	0.359	0.046	7.762	0	0.023	0.052	0.44	0.66	0.05	0.17	0.293	0.77
情绪					0.365	0.048	7.589	0	-0.255	0.153	-1.674	0.095
environment*情绪									0.344	0.178	1.936	0.054
R^2	0.141				0.142				0.151			
调整R^2	0.139				0.137				0.144			
F值	F(368, 1)=60.254，p=0.000***				F(2, 365)=30.157，p=0.000***				F(3, 364)=21.505，p=0.000***			
△R^2	0.141				0.142				0.151			
因变量：热度												

注：***、**、*分别代表1%、5%、10%的显著性水平

调节效应分析表 4 的结果显示，模型 2 到模型 3 时，△F 值的 P 值为 0.000***<0.05，呈现显著性，意味着调节变量情绪对于 environment 对热度的影响会产生显著干扰。

表 5. 情绪对服务与热度的调节效应

	模型1				模型2				模型3			
	B	标准误	t	p	B	标准误	t	p	B	标准误	t	p
const	0.142	0.042	3.419	0.001	0.219	0.062	3.561	0	0.23	0.193	1.194	0.233
service	0.212	0.068	3.137	0.002	-0.089	0.053	-1.694	0.091	0.197	0.349	0.565	0.573
情绪					0.217	0.068	3.216	0.001	-0.101	0.199	-0.508	0.612
service*情绪									0.022	0.358	0.06	0.952
R^2	0.026				0.034				0.034			
调整R^2	0.024				0.028				0.026			
F值	F(368, 1)=9.838，p=0.002***				F(2, 365)=6.379，p=0.002***				F(3, 364)=4.242，p=0.006***			
△R^2	0.026				0.034				0.034			
因变量：热度												

注：***、**、*分别代表1%、5%、10%的显著性水平

调节效应分析表 5 的结果显示，模型 2 到模型 3 时，△F 值的 P 值为 0.012**<0.05，呈现显著性，意味着调节变量情绪对于 service 对热度的影响会产生显著干扰。

表 6. 情绪对产品质量与热度的调节效应

	模型1				模型2				模型3			
	B	标准误	t	p	B	标准误	t	p	B	标准误	t	p
const	0.249	0.051	4.897	0	0.301	0.057	5.299	0	0.301	0.079	3.822	0
cost	0.023	0.079	0.291	0.771	-0.128	0.064	-2.019	0.044	0.125	0.221	0.564	0.573
情绪					0.125	0.093	1.341	0.181	-0.128	0.107	-1.194	0.233
cost*情绪									0	0.251	0.002	0.999
R^2	0				0.011				0.011			
调整R^2	-0.003				0.006				0.003			
F值	F(368, 1)=0.085，p=0.771				F(2, 365)=2.08，p=0.126				F(3, 364)=1.383，p=0.248			
△R^2	0				0.011				0.011			
因变量：热度												

注：***、**、*分别代表1%、5%、10%的显著性水平

调节效应分析表 6 的结果显示，基于交互项 cost*情绪,显著性 P 值为 0.999，模型 3 的交互项没有呈现出显著性，意味着调节变量情绪对于 cost 对热度的影响，不会产生显著干扰。

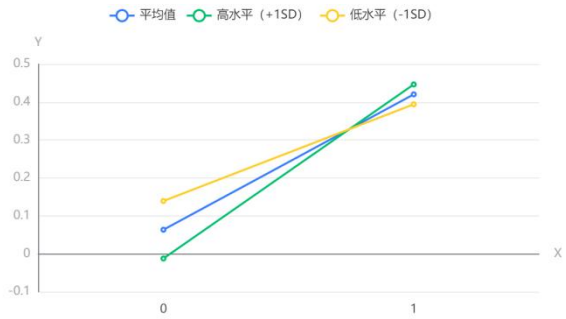
表 7. 情绪对成本与热度的调节效应

	模型1				模型2				模型3			
	B	标准误	t	p	B	标准误	t	p	B	标准误	t	p
const	0.239	0.029	8.336	0	0.312	0.056	5.535	0	0.109	0.11	0.991	0.322
quality	0.047	0.046	1.015	0.311	-0.08	0.053	-1.504	0.134	0.404	0.173	2.33	0.02
情绪					0.045	0.046	0.969	0.333	0.14	0.115	1.21	0.227
quality*情绪									-0.391	0.182	-2.149	0.032
R^2	0.003				0.009				0.021			
调整R^2	0				0.004				0.013			
F值	F(368, 1)=1.03，p=0.311				F(2, 365)=1.647，p=0.194				F(3, 364)=2.648，p=0.049**			
△R^2	0.003				0.009				0.021			
因变量：热度												

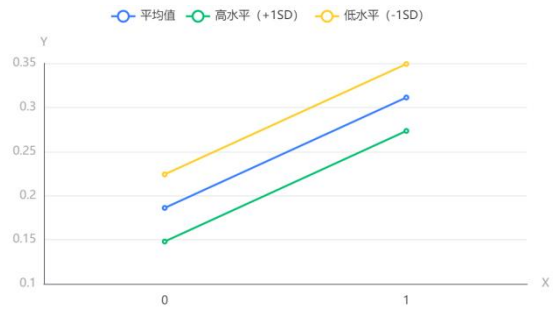
注：***、**、*分别代表1%、5%、10%的显著性水平

从表 7 中可以发现，模型 2 到模型 3 时，△F 值的 P 值为 0.013**<0.05，呈现显著性；同时基于交互项 quality*情绪，显著性 P 值为 0.032**<0.05，模型 3 的交互项呈现出显著性；意味着调节变量情绪对于 quality 对热度的影响会产生显著干扰。

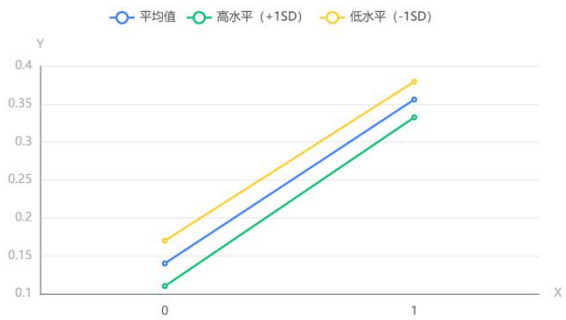
我们也绘制了简单斜率图更清晰地描述四种因素的影响程度：



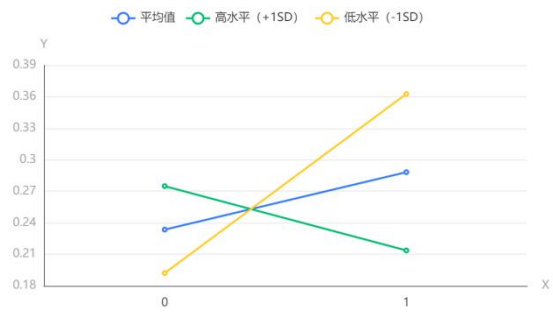
(1). 环境



(2). 成本



(3). 产品质量



(4). 服务

图 17. 情绪对多个维度的调节效应简单斜率图

从图中我们可以看到环境和服务两种因素受到情绪影响最强烈，因为斜率变化非常大。而成本因素几乎是平行线，说明基本不受情绪因素调节。产品质量因素虽然交互项结果反映也会受到情绪调节效应，但调节效应作用并不大。在环境评分高于 0.7 左右时情绪开始起到明显调节作用，而服务的临界值在大约 0.4 左右，说明受到情绪作用更早而且效果更剧烈，提高产品热度并非大力投入产品本身，而是在于提高服务水平。

最终经过结构方程模型，我们的调节效应模型图形如图 18 所示：

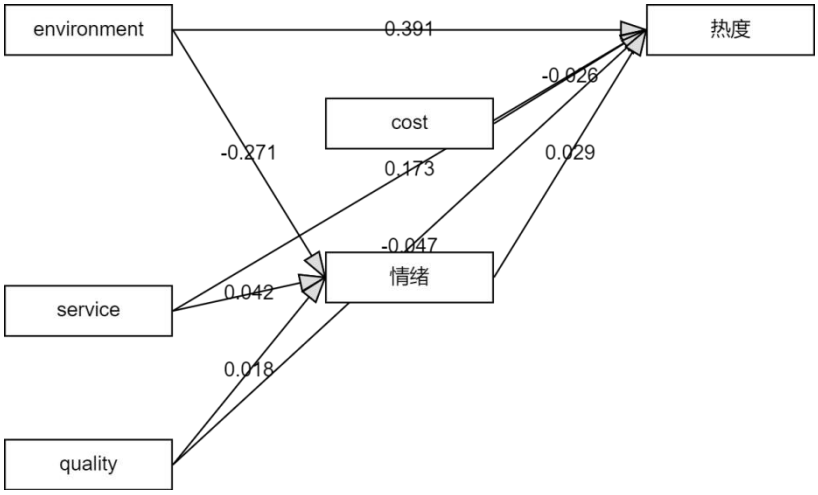


图 18. 不同变量对热度的路径分析图

对这一路径进行了 SARGAN 内生性检验，以情绪作为内生变量，分别控制变量以服务、成本、质量和环境四项因素进行了内生性检验发现，情绪变量在服务和环境两个变量的作用中具有内生性而在质量和成本两个变量的作用中并没有内生性与反向因果关系。说明菜品的服务和环境与情绪之间是复杂的作用关系，情绪会影响到环境与服务的评分，而反过来环境和服务也会影响情绪。经分析，我们认为内生性产生的主要原因在于笼统进行评价而未有考虑产品类别，我们引入产品类别作为虚拟变量以后，内生性问题得到了有效解决，故我们认为这一评价模型仍然具有较高价值。

5.3 问题三的模型

5.3.1 Apriori 算法和 PCY 算法的求解

两种算法都求解得到了 522 个频繁 1 项集，628 个频繁二项集，1375 个频繁 3 项集和 6649 条关联规则。但就时间层而言，apirori 算法花费了 112 秒但 PCY 算法仅花费了 86 秒，在时间上得到了大大优化。

频繁 1 项集的平均置信度为 0.036，频繁 2 项集为 0.02，3 项集为 0.02，这一结果反映频繁 2 项集和频繁 3 项集是相近的置信度。我们将频繁 2 项集产生的关联规则记录为两个实体产品产生了关联，将结果保存到 result3.csv 中。1 项产品关联 1 项产品的平均置信度为 0.78，而 1 项产品关联两项产品的平均置信度为 0.56，两项产品关联 1 项产品的平均置信度为 0.65，多项产品关联多项产品的平均置信度为 0.62。这一系列的模式都反映了一对一的关联模式更为可信的，但同

时也存在非一对一映射的关联规则。

5.3.2 其他关联规则分析

我们从其他三种不同的关联模式中分别取例子进行解释：

1. 一项产品关联多项产品：例如{" '饼'"} => {" '半生缘'", " '白糖罂'", " '菠萝蜜'"}: 0.7，此类关联规则说明购买产品“饼”可能关联的其他产品，并不意味着购买这一件产品一定会把其他产品捆绑销售，可能是由于二者之间在文化上存在关联或者本身存在一些特殊因素容易共现。

2. 多项产品关联一项产品：例如{" '宅'", " '饺子'", " '菠萝蜜'"} => {" '贡园'"}: 1.0，这一类关联规则则是在多种产品共现的基础上进行推断，推断出有可能关联的其他产品。

3. 多项产品关联多项产品：例如{" '油十肥佬鸭粥'", " '白粥'"} => {" '酒饮'", " '红旗督导'"}: 0.769，这一类关联较为复杂，是在组合共现以后可能产生关联的组合进行列举。

我们提取大类的关联模式后共挖掘出 25 种大类关联模式，发现平均关联度最高的几类关联模式都与文创产生了关联，这是很有意思的现象。说明文创产品能够有力促进当地旅游的发展。由于关联模式过多，不便在下文知识图谱中展示。这里将不同关联模式之间单独抽出作热力图如图 19 所示：

图 19. 不同关联模式之间的平均关联度

有趣的是，这 25 种关联模式刚好是这五类产品之间的关联。民宿之间的关联程度为 1 是因为关联模式的数量较少且都高，是否具备普遍性并不明确。而较高的还有民宿和特色餐饮之间，乡村旅游与文创之间，可以说，文创很明显地成为了关联关系中重要的影响因素。

5.4 问题四的模型

我们根据前面几个任务的结果构建了图数据库模型，基于 Neo4j 数据库逐层将“茂名旅游”、七大类产品和各自排名前 25 的产品作为三类不同级别的节点引入数据库中。由于大类之间的关联关系过多不易展示，我们并未在知识图谱中进行绘制。

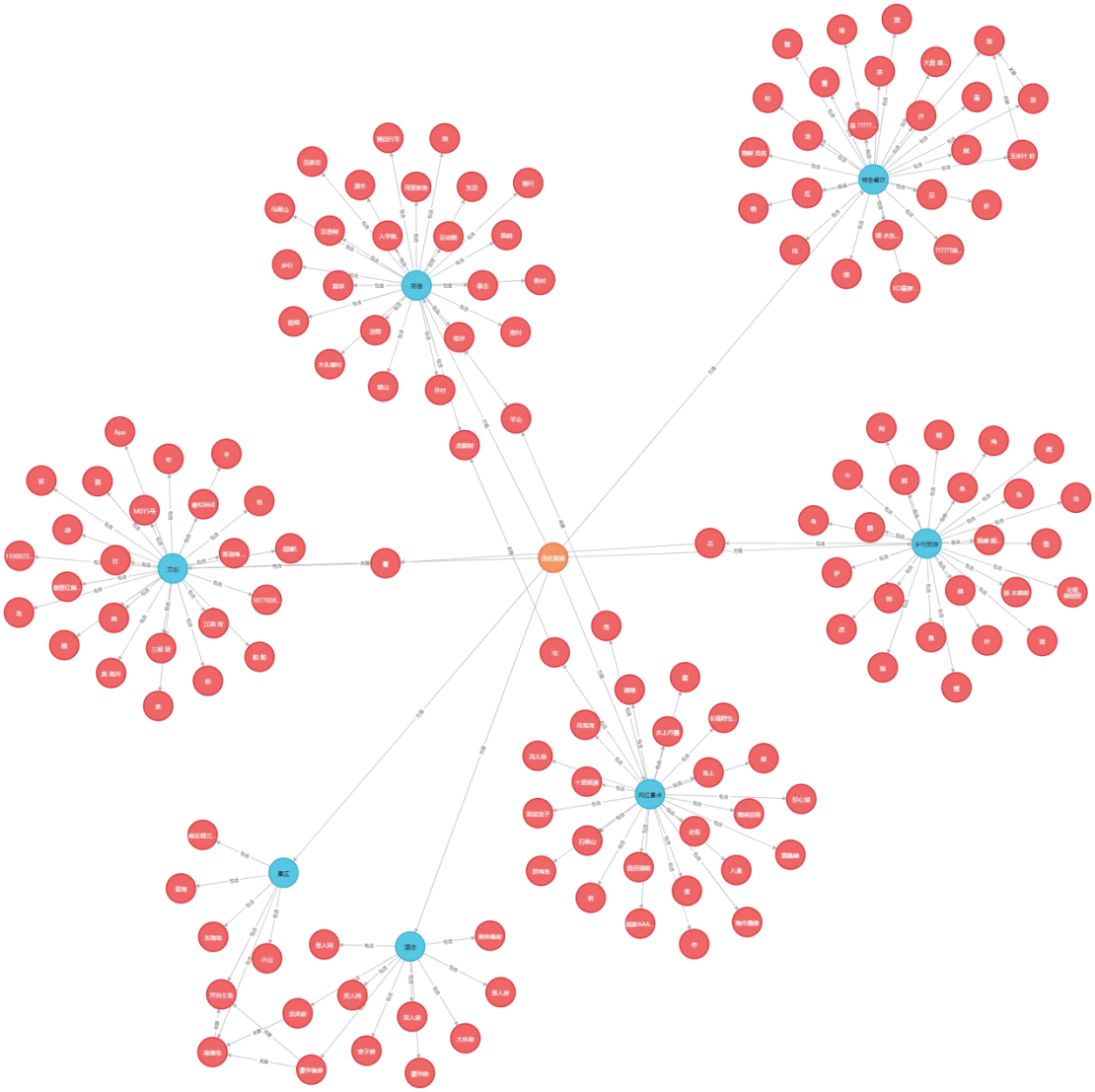


图 20. 2018-2019 茂名市本地旅游知识图谱

图 20 为 2018-2019 年茂名市旅游知识图谱，我们提取了 2018-2019 年的产品和高频产品之间的关联模式。从图 20 中可以看出，景区和酒店之间关联是比较密切的。这也与独特的自然风光和旅游服务有关。其中，东海岛、乔治王岛和南澳岛这三个岛屿的风光是最佳的，也和酒店相关服务联系非常紧密，因地制宜发展合适的旅游服务。其他诸如“玉米汁”、“虾”等菜品热度不仅高，之间联系也比较紧密，可以作为套餐推出。

在疫情来临前，景区和酒店之间相辅相成的模式为当地旅游业贡献了不小的热度，而因地制宜发展的特色餐饮产品、文创产品则起到大力的辅助作用，不同产品之间达成一个整体，使得当地旅游业更加兴旺。

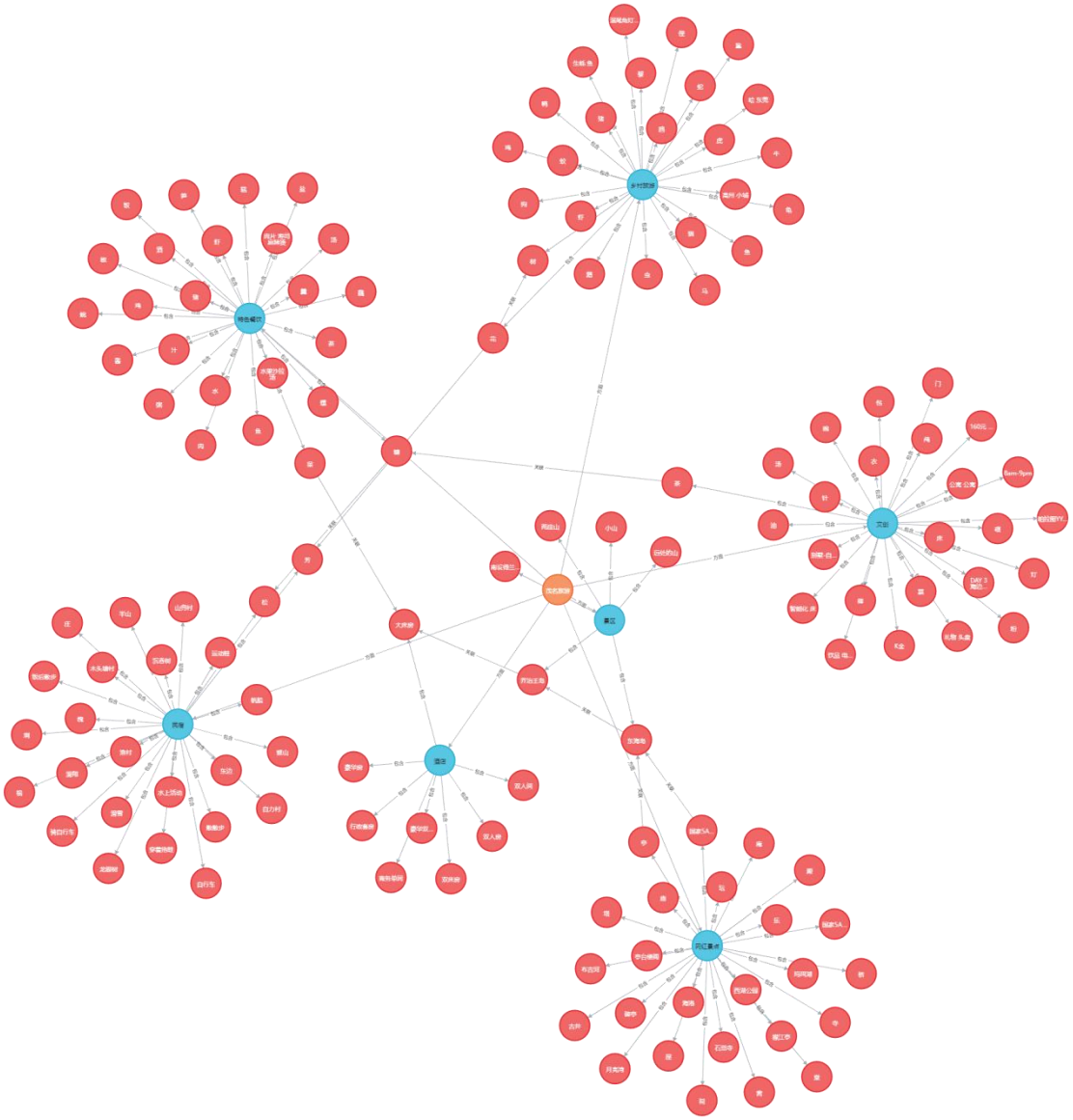


图 21. 2020-2021 茂名市本地旅游知识图谱

图 21 为 2020-2021 年茂名市旅游知识图谱，我们提取了 2018-2019 年的产品和高频产品之间的关联模式。从图 21 中可以看到，2020-2021 年产品之间的关联很显然比 2018-2019 更为复杂。其中，景区与其他产品的关联是种类最多的，包括特色餐饮、酒店、民宿等都能与景区产生关联。而这些景区中最热门的仍然是乔治王岛、东海岛和南澳岛，说明这三个地区是旅游热门景点，加强与其关联的文创、特色餐饮发展更有利于发展当地旅游业。

可以看到，由于疫情的加剧，形如酒店、民宿、特色餐饮等产品的热度普遍不如 2018-2019，但文创产品等非物质性产品却反而得到了提高，这可以看到虽然疫情对实体产业例如餐饮业服务业造成了巨大打击的同时也促进了旅游文化的转型，而景区的结构则随着时间的变化更加错综复杂，与更多种类的产品实体产生了关联。

图谱数据可以很明显揭示一些变化，我们也将会以这些变化模式撰写下面的建议信：

5.4.3 建议信

我们基于以上结论撰写了如下建议信：

==========建议信==========

茂山区旅游主管部门：

您好！在近年来新冠疫情防控常态化的背景下，游客的消费方式和消费热点发生了改变，旅游业也发生了很大的变化。经过我们的研究和分析，对于旅游业的发展，我们有如下建议：

首先，作为一个旅游区，想要吸引游客，就要有自己的特色。茂山的一个重要特色在于这里有丰富的自然景观，秀丽的自然风光可以让游客体会到亲近自然的独特乐趣。因此可以据此因地制宜，合理利用自身的旅游资源，打造自己的独特品牌。围绕自然景观和自然旅游的主题，建设度假村、主题公园、博物馆等，建立独特的旅游品牌。同时，可以大力发展乡村旅游，建设乡村生活体验点，让游客回归自然，真切地体验到当地的生活，感受当地的民俗。

第二，发展旅游业需要建立完善的服务区。良好的服务区可以改善游客的出游体验，提升游客旅行满意度。服务区包括酒店、民宿、餐饮等多个方面。对于酒店和民宿而言，首要的要求是注重服务和卫生，良好舒适的居住环境可以给游客带来更好的体验。同时，性价比也是影响游客旅游住宿选择的重要的因素。在餐饮方面，在注重菜品和口味的同时也要注重环境和服务。也可以与当地特色美食结合，打造独特品牌。服务区的位置也是游客关注的方面，为了让游客有更好的旅游体验，可以将景区和酒店餐饮相结合，让游客可以在休息的同时可以欣赏美丽的自然风光，也可以减轻游客在酒店和景区间路途的劳累。

第三，发展文创产品。文化是旅游业发展的竞争力，一个有文化底蕴的地方会让人印象深刻，游客在选择旅游目的地时会更注重一个地方的文化底蕴。而文创产品是可以让“文化”伴随游客的方式，游客购买文创产品作为纪念品，会在其心中留下深刻印象。文创产品可以丰富游客旅游体验在经济、用户体验、品牌传播多方面起到推进作用。一方面可以进一步宣传地方特色、推动品牌传播、增加品牌影响力，另一方面可以推动当地经济的发展。

第四，在新冠疫情防控常态化的环境下，游客对防疫政策格外关注，因此当地应明确防疫政策，景区和酒店等也要明确要求，让游客有充足的准备。对于酒店、民宿、餐馆也要加强培训、管理和指导，完善防疫措施。同时，要对于疫情导致的突发状况做好应急预案，细化各种措施，以应对疫情突然发生，做到从容应对而不能慌乱无措，可以合理安排好游客，有充足的物资，尽量将相关因素带来的负面影响降至最低。

第五，在疫情的环境下，游客出行受阻，有时不能进行亲身实地的旅游出行。因此可以尝试发展数字化旅游，建立数字博物馆、网络导游等相关设施，扩展新的数字创收增长点。一方面，丰富了游客的旅游形式和旅游体验，让游客们可以足不出户进行云旅游、云体验。另一方面，这也是一种宣传的方式，通过网络生动形象全面地展现出茂山的特色风土人情，可以增加热度，吸引更多的游客。

总而言之，结合当下防疫常态化背景，发展旅游业首先需要相关旅游设施和服务设施是完善，同时做好防疫应急预案，在此基础上做出一定的创新，使本地旅游业健康蓬勃发展！

六. 模型的评价与分析

6.1 模型的优点

我们经过多种数据挖掘手段有效解决了业务需求。经过相对严密的数据驱动分析，我们认为，我们的模型具有以下优点：

1. 同一问题使用不同方法进行对比，保证能够选择到最优模型进行预测。在对比的过程中我们也综合衡量了多项技术指标，在保证模型准确性的同时提高模型运行效率。

2. 灵活选用不同工具，在 python 的基础上选择使用 SPSSPRO 进行复杂的路径分析、统计检验与结构方程，又合理选择 Neo4j 数据库进行知识图谱的构建与可视化，手段较为灵活，且降低了开发成本。

3. 数据安全可靠，Neo4j 作为一种图数据库其数据安全性有较强保证，不容易受到攻击，以此类方式存储数据能够对海量数据进行安全、可靠的存储与分析。

4. 我们还在原有基础上对热度的影响因素以及作用机理进行了进一步探究，使得结果更具有可信度，也为热度提升提供了具有一定参考价值的意见。

6.2 模型的缺点

客观公正地说，我们也认为我们的模型具有以下几点局限性：

1. 数据体量过小且缺乏标注，我们的标注方法比较简单，按照我们自定义的标注模型是高效的，但标注可能具有一定偏差。

2. 我们对多种模态的数据进行了混合分析，并未考虑到文本对象之间的差异性。如果在问题中将不同评论对象作为一个新的变量引入模型可能会更好。

3. 我们抽取了多种不同的维度进行评价，但子句抽取可能不一定准确，加上 SnowNLP 本身情感给分的误差性，可能导致结果有一定偏差。

6.3 模型的总结和改进

我们通过自然语言处理和知识图谱的手段对所提出的问题建立合理的模型并提出自己的解决方案。综合运用了文本分类、情感分析、命名实体识别、关联模型、评价模型、结构方程和知识图谱等多种数据挖掘任务方法，灵活运用 python、SPSSPRO、Neo4j 数据库等多种工具，对所给定的问题进行了合理解答。通过数据驱动和知识驱动的方法与视角对给定问题提出了深入探究和解决方案，具有一定参考价值。

而对于模型尚且存在的一些局限性问题，我们认为如果模型能够对不同的语料进行对比分析，在大规模有标注旅游文本上训练以后再进行迁移学习可能会更好。

参考文献

[1]施聪莺，徐朝军，杨晓江. TFIDF 算法研究综述[J]. 计算机应用, 2009, 29（B06）:167-170.

[2]Cavnar W B， Trenkle J M . N-Gram-Based Text Categorization. 2001.

[3]Mikolov T， Chen K， Corrado G，et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[4]Pennington J， Socher R， Manning C . Glove: Global Vectors for Word Representation[C]// Conference on Empirical Methods in Natural Language Processing. 2014.

[5] Suykens J， Vandewalle J . Least Squares Support Vector Machine Classifiers[J]. Neural Processing Letters, 1999, 9（3）:293-300.

[6]Mccallum A， Nigam K . A comparison of event models for Naive Bayes text classification[J]. IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION, 1998:41--48.

[7]彭君睿. 面向文本分类的特征提取算法研究[D]. 北京邮电大学, 2014.

[8]Kim Y . Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.

[9]Zhang Y， Wallace B . A Sensitivity Analysis of （and Practitioners' Guide to） Convolutional Neural Networks for Sentence Classification[J]. Computer Science, 2015.

[10] Liu P， Qiu X， Huang X . Recurrent Neural Network for Text Classification with Multi-Task Learning[J]. AAAI Press, 2016.

[11]Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification[C]//Twenty-ninth AAAI conference on artificial intelligence. 2015.

[12]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[13]Yang Z， Yang D， Dyer C , et al. Hierarchical Attention Networks for Document Classification[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.

[14]Peng Z， Wei S， Tian J , et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics （Volume 2: Short Papers）. 2016.

[15]Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 562-570.

[16]He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[17]Jozefowicz R， Vinyals O， Schuster M , et al. Exploring the Limits of Language Modeling[J]. 2016.

[18]Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving Language Understanding by Generative Pre-Training[J]. arXiv, 2018.
[19] Devlin J， Chang M W， Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
[20]Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.
[21]Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[J]. Advances in neural information processing systems, 2013, 26.
[22]孙镇,王惠临.命名实体识别研究进展综述[J].现代图书情报技术,2010（06）:42-47.
[23]万静,涂喆,冯晓.基于条件随机场的医药领域症状信息抽取[J].北京化工大学学报（自然科学版）,2016,43（01）:98-103.
[24]姜文志,顾佼佼,丛林虎.CRF 与规则相结合的军事命名实体识别研究[J].指挥控制与仿真,2011,33（04）:13-15.
[25]O Cappé， Moulines E , T Rydén. Inference in Hidden Markov Models[J]. Technometrics, 2006, 48（4）:574-575.
[26]何炎祥,罗楚威,胡彬尧.基于 CRF 和规则相结合的地理命名实体识别方法[J].计算机应用与软件,2015,32（01）:179-185+202.
[27]Lafferty J， Mccallum A， Pereira F . Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]// Proc. 18th International Conf. on Machine Learning. 2001.
[28]Imry Y， Ma S K . Random-Field Instability of the Ordered State of Continuous Symmetry[J]. Physical Review Letters, 1975, 35（21）:1399-1401.
[29]Graves A, Jaitly N, Mohamed A. Hybrid speech recognition with deep bidirectional LSTM[C]//2013 IEEE workshop on automatic speech recognition and understanding. IEEE, 2013: 273-278.
[30]Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
[31]Schuster M, Paliwal K K. Bidirectional recurrent neural networks[J]. IEEE transactions on Signal Processing, 1997, 45(11): 2673-2681.
[32] Wang, Xuan, Xu, et al. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN[J]. Expert Systems with Application, 2017.
[33]陈伟,吴友政,陈文亮,张民.基于 BiLSTM-CRF 的关键词自动抽取[J].计算机科学,2018,45（S1）:91-96+113.
[34]许力, 李建华. 基于 BERT 和 BiLSTM-CRF 的生物医学命名实体识别[J]. 计算机工程与科学, 2021, 43(10):7.
[35]谢腾, 杨俊安, 刘辉. 基于 BERT-BiLSTM-CRF 模型的中文实体识别[J]. 计算机系统应用, 2020, 29(7):8.
[36]R. Agrawal, T. Imielinski, and A. Swami, “Mining associations between sets of items in massive databases,”Proc. ACM SIGMOD Intl. Conf. On Management of Data, pp. 207–216, 1993.
[37]J.S. Park, M.-S. Chen, and P.S. Yu, “An effective hash-based algorithm for mining association rules,” Proc. ACM SIGMOD Intl. Conf. on Management of Data, pp. 175–186, 1995.

[38]杨炳儒. 知识工程与知识发现[M]. 冶金工业出版社, 2000.
[39] Liu Z , Han X . Deep Learning in Knowledge Graph[M]. 2018.
[40] Partner J , Vukotic A , Watt N . Neo4j in Action[J]. Pearson Schweiz Ag, 2014.
[41]刘雪梅. 上市公司治理结构与盈余管理关系研究[D].华北电力大学（北京）,2008.
[42]Chen C T . Extensions of the TOPSIS for group decision-making under fuzzy environment[J]. Fuzzy Sets & Systems, 2000, 114(1):1-9.
[43]程启月.评测指标权重确定的结构熵权法[J].系统工程理论与实践,2010,30(07):1225-1228.
[44]谢宏,程浩忠,牛东晓.基于信息熵的粗糙集连续属性离散化算法[J].计算机学报,2005(09):1570-1574.
[45]Zhang H, Liu L, Jiang H, et al. Texsmart: A text understanding system for fine-grained ner and enhanced semantic analysis[J]. arXiv preprint arXiv:2012.15639, 2020.
[46]https://download.csdn.net/download/qq_38295507/11110576?utm_medium=distribute.pc_aggpage_search_result.none-task-download-2~aggregatepage~first_rank_ecpm_v1~rank_v31_ecpm-1-11110576.pc_agg_new_rank&utm_term=%E7%BE%8E%E5%9B%A2%E8%AF%84%E8%AE%BA%E6%95%B0%E6%8D%AE%E9%9B%86&spm=1000.2123.3001.4430

附录

环境：Python 3.7.6 + Pytorch 1.10.2 + tensorflow 2.2.0 + CUDA 10.2 + Neo4j 4.4.6 Windows 10 系统，GPU 为 NVIDIA GTX1650， Intel i7 型 CPU

知识图谱构建代码：

```
# -*- coding: utf-8 -*-
"""
Created on Thu Apr 21 21:48:54 2022
@author: Mashituo
"""
import pandas as pd
from py2neo import Graph, Node, Relationship, NodeMatcher
graph=Graph('http://localhost:7474',auth=('*****','*********'))
data=pd.read_csv("newdf2_year.csv")
data=data[data['years']=='2018-2019']
Main=Node("Main", name='茂名旅游')
_type=data['特色产品'].unique()
nodes=[]
for _t in _type:
    Types=Node("Type",name=_t)
    entity = Relationship(Main,"方面", Types)
    graph.create(entity)
    subdf=data[data['特色产品']==_t]
    subdf=subdf.sort_values("热度").reset_index(drop=True)
    subdf=subdf[['产品名称','特色产品']].drop_duplicates()[0:25]
    subnodes=[]
    for product in subdf['产品名称']:
        Prod=Node("Product",name=product)
        subnodes.append(Prod)
        subentity = Relationship(Types, "包含", Prod)
        graph.create(subentity)
    nodes+=subnodes[0:3]
pairdata=pd.read_csv("pcy_year.csv")
matcher = NodeMatcher(graph)
for i in range(len(pairdata)):
    try:
        pro1=list(matcher.match("Product").where(name=pairdata['item1'][i]))[0]
        pro2=list(matcher.match("Product").where(name=pairdata['item2'][i]))[0]
        crossentity = Relationship(pro1,"关联", pro2)
        graph.create(crossentity)
    except:
        Pass
# graph.delete_all()
```