第十届泰迪杯数据挖掘挑战赛云颁奖暨数字人才发展交流会

第十届泰迪杯C题作品点评

邓明华 dengmh@math.pku.edu.cn
北京大学数学科学学院

2022年7月23日

提纲

- 题目与分析
- 论文总体情况
- 评阅中发现的问题
- 优秀论文亮点展示

问题背景

- 随着互联网和自媒体的繁荣，文本形式的在线旅游（Online Travel Agency，OTA）和游客的用户生成内容（User Generated Content，UGC）数据成为了解旅游市场现状的重要信息来源。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。
- 为此本赛题提出本地旅游图谱这一概念，它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。本地旅游图谱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。
- 在近年来新冠疫情常态化防控的背景下，我国游客的旅游消费方式已经发生明显的转变。在出境游停滞，跨省游时常因为零散疫情的影响被叫停的情况下，中长程旅游受到非常大的冲击，游客更多选择短程旅游，本地周边游规模暴涨迎来了风口。疫情防控常态化背景下研究分析游客消费需求行为的变化，对于旅游企业产品供给、资源优化配置以及市场持续开拓具有长远而积极的作用。本赛题提供收集自互联网公开渠道的 2018 年至 2021年广东省茂名市的 OTA 和 UGC 数据，期待参赛者采用自然语言处理等数据挖掘方法通过建立本地旅游图谱的方式来分析新冠疫情时期该市周边游的发展

问题1：微信公众号文章分类

- 构建文本分类模型，对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。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1分析

- 分类问题，但缺乏标注；
- 题目给出了旅游相关的部分关键词
- 思路一：首先利用题目给出的关键词进行人工标注，然后学习一般分类模型；甚至可以引入主动学习的思想，对难以区分的词条挑出来进一步扩充学习学习集。
- 思路二：无监督聚类，根据聚类结果进行分类；理想情况是：如果某个子类含有题目相关的关键词，则可以判断成相关。

问题2：周边游产品热度分析

- 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保存。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。

问题2分析

- 提取旅游产品提取：景区、酒店、餐饮中的旅游产品可能比较明确，但游记和旅游公众号的旅游产品需要通过实体命名识别方法进行提取。

- 建立多维度热度评价模型，模型的合理性。文本情感分析可以是热度分析的出发点。

问题3：本地旅游图谱构建与分析

- 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，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。并且力争挖掘旅游产品间隐含的关联模式并进行解释。

问题3分析

- 关联规则发现，关键词共现等方法挖掘旅游产品之间的关联；

- 游记或者公众号中的时间先后、一条龙服务等可能提供一定的隐含关联，

问题4：疫情前后旅游产品需求的变化分析

- 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。
- 分析：该问题主要是一个旅游产品需求的疫情前后对比分析，政策建议需要“言之有数”

论文总体情况

- 随着泰迪杯挑战赛的深入，大家对文本处理的基本流程还是比较熟悉，掌握了关键词提取、文本向量化、以及多种端到端的深度学习方法；
- 比较熟悉常用的机器学习方法，熟练使用Python机器学习包完成分类任务；

典型求解方法

- 问题1：TF-IDF或者Word2vec向量化，机器学习方法(如NB, RF, SVM等等)进行分类；
- 问题2：TextRank, LDA, CRF等实体命名识别方法；SnowNLP情感识别；AHP, TOPSIS或者模糊评价；
- 问题3：Apriori, FP-Growth等关联规则发现，灰色关联分析；
- 词云比较对比差别；

评阅中发现的问题

- 问题1：虽然是分类问题，但很多同学没有把笔墨用在训练集合的构造上，而大部分笔墨放在分类方法的数学模型介绍上，抄录了一大堆公式；缺少多种方法效果对比；
- 问题2：很大部分同学没有游记和微信公众号进行旅游产品提取；综合评价三板斧(AHP, TOPSIS, 模糊评价)
- 问题3：隐式关联分析很少；
- 问题4：疫情前后对比比较粗糙，政策建议比较空泛。

优秀论文亮点展示

- 001C和464C各有特色
- 001C论文引入了对偶对比学习技术，解决了旅游产品相关训练样本数据缺乏的问题；采用图神经网络进行链路预测，抽取出旅游产品间的隐含关系；
- 464C获奖论文能够充分考虑问题特征，进行模型选取及优化；调用百度API可视化效果挺不错。

001C亮点(1)

- 问题1文本分类
  - 通过爬虫技术获取旅游类文本并结合THUCNews新闻文本分类数据集，构建旅游文本分类训练集；
  - 文本摘要减少输入文本长度(BiGRU>TextRank)；
  - 基于RoBERTa-BiGRU-Attention 融合模型进行文本分类；
  - 对偶对比学习，标签启发式增强样本

分类效果

表 2- 4 典型模型的文本分类结果评估对比

模型	Loss	Acc(%)
SVM	0.4481	90.921
Bayes	0.4011	91.612
RNN	0.2937	91.632
LSTM	0.2455	91.894
Text CNN	0.2234	91.793
Text RNN	0.2256	91.833
BERT	0.1940	92.534
Roberta	0.1672	92.976
Roberta-BiGRU	0.1340	93.112
Roberta-BiGRU-MA	0.0913	93.572
Roberta-DuaCL	0.00186	96.900

001C亮点(2)

- 问题3隐含关系抽取
  - 先利用改进**Apriori** 算法进行关联分析，并抽取(实体，关系，实体)三元组，相较于传统算法效率更高；
  - 构建了**GNNLP**模型，以抽取出的三元组为基础建图，并利用图神经网络进行链路预测，对隐关联进行挖掘；
  - 测试结果表明，利用GNNLP 模型挖掘出的隐含关联关系较普通单一模型效果平均提高了11.76%。

强关联与隐含关联

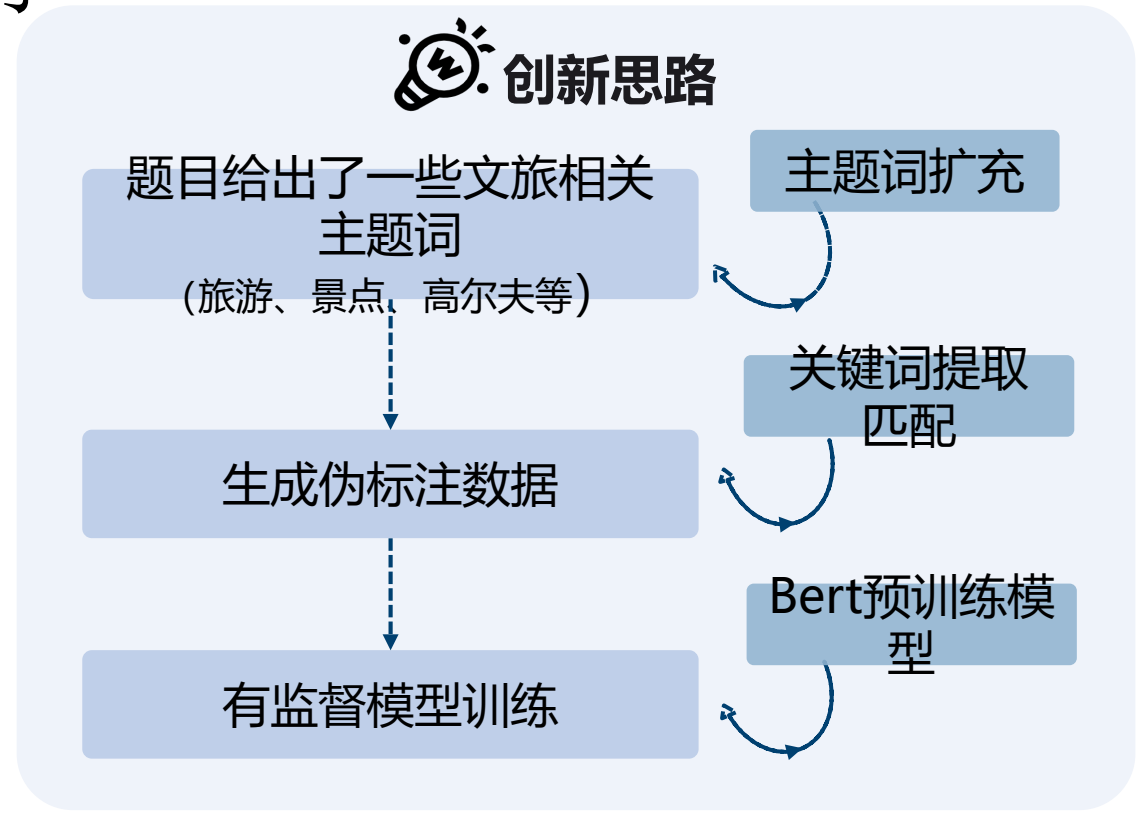
产品1	产品2	关联度	关联类型
水东芥	捞粉	0.8	餐饮—餐饮
鸡心黄皮	菠萝蜜	1.0	餐饮—餐饮
街头牛排	温德姆酒店	0.6666	餐饮—酒店
食惯嘴	御水古温泉酒店	0.5	餐饮—酒店
鸡冠石	放鸡岛	0.6666	景点—景点
大角湾	十里银滩	0.75	景点—景点

产品1	产品2	关联关系
海陵岛	海景湾大酒店	景点—酒店
何记清补凉	茂名博物馆	餐饮—景点
希亚酒店	聚佳购物广场	酒店—景点
如家酒店	浪漫海岸	酒店—景点
龙腾国际	黄潮牛庄	酒店—餐饮
喜来登酒店	槐花宴	酒店—餐饮

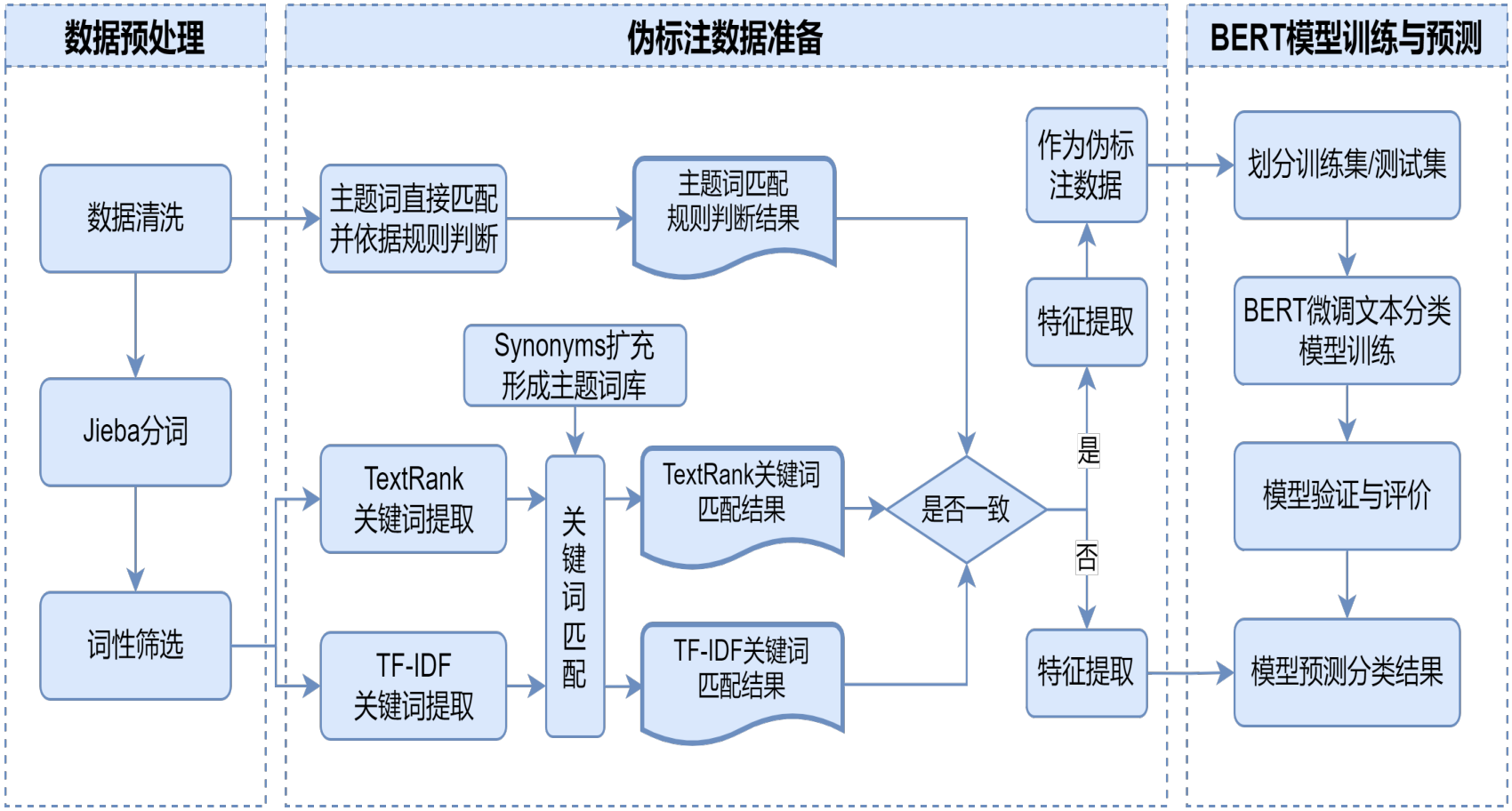
海陵岛和海景湾大酒店通过上层概念捞粉建立关系

464C亮点

- 问题1求解方法：扩充词库，构造伪标签进行学习



464C亮点



注:

1. 本文提及的主题词是指题目中所指跟文旅相关与否的主题词；关键词是指文章提取的中心词。
2. 特征提取的做法是：选取TextRank和TF-IDF形成的前50个关键词与文章标题的核心词去重后作为文章特征。

将依赖人工标注的数据转化为自动标注的数据，并且模型准确率高达97%。

三个标签一致：4945条

文章ID	公众号标题	正文	关键词提取(50)	label (TFIDF)	label (TextRank)	label (RuleBased)
1001	2018，对自己好一点	2017的旅程已经结束2018	['身体', '本钱', '帷幕', '自律',	0	0	0
1002	春节机票预订有窍门	距离春节还有一个多月的时	['汽车', '机票', '飞机', '时间',	0	0	0
1003	冬日旅游知多D	960万平方公里的祖国大地	['雪盲', '暖手', '媒体报道', '将	1	1	1
1004	2018冬季暖心之旅	长按二维码，关注我们中心	['市区', '心', '电白', '联系人',	0	0	0
1005	关于粤K27618号大客		['大客车', '喷火', '排气管', '设	0	0	0
1006	2018年全国旅游工作会	1月8日全国旅游工作会议在	['质量', '导游', '游客', '政策',	1	1	1
.....	.....	.....	.....	.....	.....	.....
7279	划重点！一图读懂高州	10月29日，中国共产党高州	['委员会', '读懂', '代表', '成绩	0	0	0
7282	报名截止时间延期！龙	由市创文巩卫办、市教育局	['小视频', '延期', '视频', '学生	0	0	0
7283	党史百年天天读·10月	重要论述1984年10月31日	['汽车', '政务院', '党史', '市场	0	0	0
7285	报名截止时间延期！龙	由市创文巩卫办、市教育局	['小视频', '延期', '视频', '学生	0	0	0
7286	党史百年天天读·10月	重要论述1984年10月31日	['汽车', '政务院', '党史', '市场	0	0	0

三个标签不一致：1341条

文章ID	公众号标题	正文	关键词提取(50)	label (TFIDF)	label (TextRank)	label (RuleBased)
1010	春节邀亲朋好友自驾游	询电话：0668-2265726	['市区', '交委', '电话', '地址',	1	0	1
1011	春节长线近游：湘西、	咨询电话：0668-226572	['市区', '长线', '游', '交委', '场	1	0	0
1012	"幸福爸妈"出游预告：	以下专列从湛江始发，出	['爸妈', '汽车站', '先生', '车辆	1	1	0
1016	春节去哪里，手指点这	春节出行，桂林赏花，茂名	['去', '分部', '电白', '总部', '作	1	0	0
1017	三八妇女节出游预告	多彩张家界以下专列从湛江	['精彩', '预告', '收客', '电白',	1	1	0
1019	三八妇女节出游预告	多彩张家界以下专列从湛江	['精彩', '预告', '收客', '电白',	1	1	0
.....	.....	.....	.....	.....	.....	.....
7276	关注！又一地发现阳性	30日下午又一地通报发现	['麦积', '社棠', '学生', '麦积区	1	1	0
7278	关注！又一地发现阳性	30日下午又一地通报发现	['麦积', '社棠', '学生', '麦积区	1	1	0
7280	关注！又一地发现阳性	30日下午又一地通报发现	['麦积', '社棠', '学生', '麦积区	1	1	0
7281	赞！茂名这两个村入选	近日农业农村部官网公布	['社会', '乡镇党委', '全国', '基	1	1	0
7284	赞！茂名这两个村入选	近日农业农村部官网公布	['社会', '乡镇党委', '全国', '基	1	1	0

Bert模型训练

Bert模型预测

模型评价

	准确率
训练集(70%)	98.23%
测试集(30%)	96.57%

预测结果

文章ID	bert_predict
1010	1
1011	1
1012	1
1016	1
1017	1
1019	1
.....	.....
7276	0
7278	0
7280	0
7281	0
7284	0

最终结果：result1.csv

文章ID	分类标签
1001	不相关
1002	不相关
1003	相关
1004	不相关
1005	不相关
1006	相关
.....	.....
7282	不相关
7283	不相关
7284	不相关
7285	不相关
7286	不相关

谢谢观看！