大数据分析与应用课程说明

Intro to Big Data Analysis and Application

张华平 副教授 博士

Email: kevinzhang@bit.edu.cn

http://www.nlpir.org/

@ICTCLAS张华平博士

大数据搜索与挖掘实验室 (BDSM@BIT)

2021-9

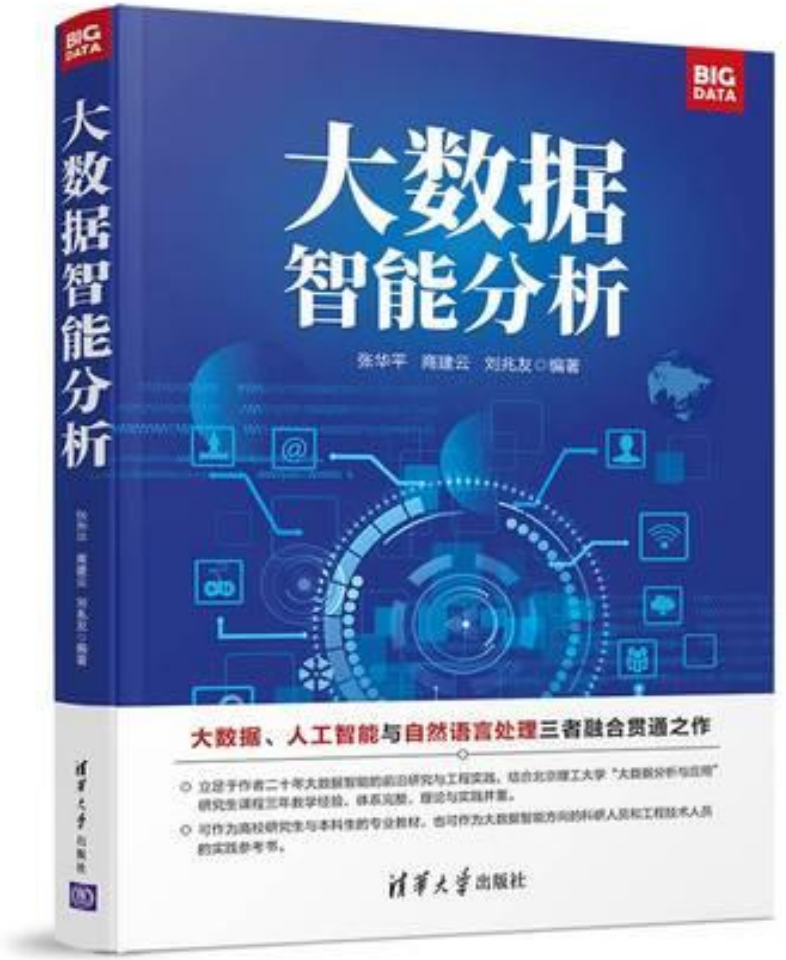
沟通渠道

- 微信群：不得发与课程无关的内容；
- 网站：http://www.nlpir.org/
- Github： https://github.com/Dr-Kevin-Zhang/Big-Data-Analysis-and-Application-Course
- 所有课程资料、同学的综述报告以及期末作业全部对外公开、B站+MOOC；
- 论证主持： 汤泽阳 雷沛可
- 摄像宣传： 李静 李育霖 张恒瑀

教材

- 张华平,商建云，刘兆友.大数据智能分析[M].北京:清华大学出版社（2019）（ISBN:978-7-302-53117-3）北理工十三五优秀教材

- 天猫 清华大学出版社旗舰店47.84；
- 参考的Baseline。



北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

讲者介绍

张华平 博士

- ICTCLAS汉语分词创立者

  创建并运营NLPIR大数据语义增强分析平台

- 北京理工大学副教授，大数据搜索与挖掘实验室主任，

  中国人工智学会多语种智能信息处理专委会秘书长

✓ 中文信息处理领域最高奖：钱伟长中文信息处理一等奖
✓ 新疆自治区科技进步二等奖
✓ 第一届ACL-SIGHAN国际汉语分词大赛
✓ 国家973汉语评测
✓ 中央网信办、中宣部、公安部等部委特聘技术顾问
✓ 国办电子政务总体组等专家

钱伟长一等奖证书

新疆科技进步二等奖

张华平教授受CCTV采访解读苹果FBI揭秘大战。

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

教育方式

兴趣第一

- 感兴趣找方法，不感兴趣找借口；
- 教育第一原则是培养对科学或者具体学科的兴趣，扼杀青年的兴趣，罪莫大焉；
- 再好的学问，以面目可憎的形象出现，年轻人也不可能接受。佛家无色无相，却幻化万象，以渡众生。

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

教育方式

知行合一

- 明 王守仁 《传习录》卷 教育家：陶行知
- 王守仁，号阳明先生，中国明代最著名的思想家、哲学家、文学家和军事家。陆王心学之集大成者，非但精通儒家、佛家、道家，而且能够统军征战，是中国历史上罕见的全能大儒。封“先儒”，奉祀孔庙东庑第58位。
- 计算机科学尤其强调知行合一。

结课成绩构成

- **平时10分**
  - 课堂考勤+互动 10分；
- **大数据智能分析技术综述报告(交付物：综述报告与PPT)：40分**
  - 最多6人一组，可自由组合，需标明分工；报告按照《计算机学报》综述报告发表要求；需要超出《大数据智能分析》
  - 评分标准： 经典综述 30%，前沿进展：20%，技术Demo 20%，美观度10%，讲解配合20%
- **大数据智能分析应用项目（交付物：代码，说明文档，演示PPT,论文） 50分**
  - 最多6人一组，可自由组合，需标明分工；
  - 可以是某项技术Demo，也可以是成熟技术的新应用；使用开源等一切资源，但不能是**简单照搬抄袭（杀无赦）**
  - 评分标准：创新与特色（40%） 应用价值（20%） 工作难度（20%）完成质量（演示与演讲）（20%）。

如何拿90+？

- 比例控制在10%；名额控制在17人
- 项目赛（最后一周）：每组10分钟演讲+演示答辩，外聘评审专家组（包括学术研究、应用部门、工业部门、投资机构）打分；
- 将署名并入选北理工教材《语义智能分析》，拟由清华大学出版社出版

多语种多模态语义分析

场景层

- 公安：警情分析；WW预测；重复报警
- 军工：情报平台；智能参谋；文书核查；会议纪要；辅助定密
- 政府/企业：投诉分析；会议纪要；公文核查；舆情监测
- 金融：合规风险；票据比对；信息抽取
- 科教：论文核查；专题平台
- 国际化：中文处理；汉语教学
- …：…；…

产品层

公共安全语义智能分析　智能文本核查　智能会议纪要　知识图谱加工厂

NLPIR语义增强分析平台（NLPIR语义分析引擎+JZSearch搜索+九眼过滤+KGB知识图谱工具等）

技术层

- 数据导入：精准采集；OCR；语音识别；DB导入；文档解析
- 多语分析：分词标注；依存文法；情感分析；语言模型；文本校对
- 信息抽取：新词发现；摘要提取；实体抽取；关键词提取；过滤与脱敏
- 文本挖掘：文档去重；相似比对；文本分类；文本聚类；话题分析
- 知识关联：全文检索；地图可视化；词向量扩展；知识生成；知识推理

基础层

- 算源：语料库；知识库；外部库
- 算法：专家启发；机器学习；终身学习
- 算力：国产芯片；国产OS；可控云

NLPIR inside

语义增强分析平台

- ✓ 具备70-80%应用功能
- ✓ 一周实现目标测试
- ✓ <100份样本冷启动
- ✓ 降低20-30%开发成本
- ✓ 提升系统智能化水平

大数据智能分析技术2021选题

- 平台层： 1. 深度学习平台：TensorFlow/PyTorch平台；
- 算法模型：2. 深度学习算法基础；3、NLP深度学习新进展；
- 数据导入：4. 爬虫；5.多格式文档解析与管理 6.语音文字识别与说话人识别；7.OCR及领域优化；8. 图像Caption-看图说话
- 语言分析：9 语言预训练模型与应用（GPT3.0; BERT;ELMo）；10.汉语分词与标注；11.句法分析; 12.情感分析; 13.多语种语言模型与处理（mBert）;14.藏语NLP；15.维语NLP；16.阿拉伯语NLP；17.机器翻译
- 信息抽取：18.新词发现; 19.关键词提取; 20.专用命名实体抽取; 21.信息过滤；22.个人隐私保护与脱敏；23.计算机图像实体检测
- 文本挖掘：24.文本分类；25.文本聚类；26.话题发现;27. 语义相似度计算；28.文本校对
- 知识关联：29.知识图谱构建；30.社交网络搜索与挖掘；31.数据融合与多模态分析；

实验案例

十大实验案例

	主题	工具	数据
1	十九大报告主题自动分析	新词，关键词分析	十九大报告
2	方文山与汪峰歌词智能对比挖掘	分词、语言模型	歌词文本
3	基于用电数据的大厦空置率预测	数据挖掘	样例数据
4	文章抄袭自动检测	关键词提取，去重	样例数据
5	微博用户画像与内容推荐	关键词提取，相似度计算	部分微博数据
6	新闻热点话题的发现	聚类	新闻数据
7	人工智能领域近三年研究创新点对比与综合	关键词、词频、摘要	AI论文题录数据
8	垃圾邮件中犯罪线索的智能发现	智能过滤	假发票等样例数据
9	产品点评情感综合判别	情感分析	京东等产品点评
10	科技文献自动分类	分类	文献数据

实验案例

30个经典作品赏析

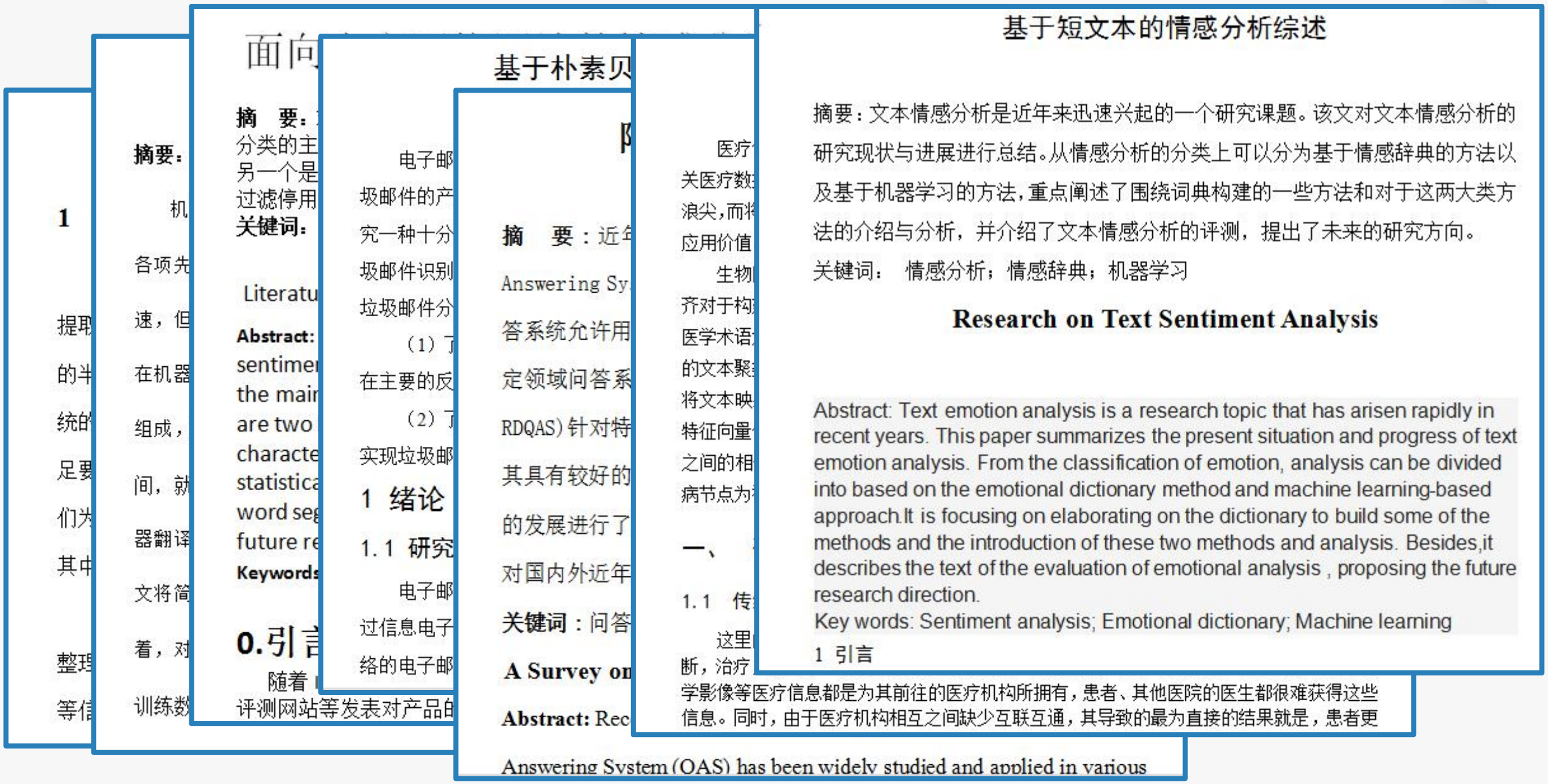
1	图像描述的智能生成	16	基于Mahout的电影推荐系统
2	基于时空推理的气象公告自动生成	17	大数据技术在医疗领域的应用
3	微博用户行为模式研究及其应用	18	面向特定领域的信息抽取与知识图谱的构建
4	微博特定群体发现模型研究	19	面向中文网络评论的情感分类研究
5	社交网络水军识别	20	基于静态图像的人物角色识别
6	跨语言图像检索系统的研究与实现	21	大数据下的医疗疾病状况分析
7	基于hadoop的垃圾邮件分类	22	基于SVM的文本情感分类研究综述与实现
8	基于LSTM模型的影评的情感倾向性分析算法实现及应用	23	交友社区中的自动匹配
9	基于SPARK的微博情绪分析	24	数据挖掘在股票预测分析上的应用
10	使用PageRank算法分析微博用户影响力	25	精准营销中用户画像挖掘
11	基于搜索日志的用户画像简易构建方案	26	基于微博的热词抽取
12	电子商务网络水军分析综述	27	神经网络机器翻译的研究与实践
13	跨语言医学术语对齐技术研究	28	学术领域问答系统的研究与实现
14	基于文本数据挖掘的商品评论情感分析	29	基于情感维量特征提取的图像情感分析
15	基于 Hadoop 的 K-Means 算法实现消费数据分析	30	基于深度学习的短文本情感分析论文综述

2020选题

- ATTA_个人语言特征消除工具
- O2O优惠券使用预测技术研究与应用
- “土拨鼠”动漫风头像生成器
- 《水浒传》文本可视化
- 大数据考研分析
- 电影推荐系统
- 丁真走红事件网络舆情分析
- 基于TextCNN模型B站疫情相关评论情感分析
- 基于bert的虚假新闻检测
- 基于出租车订单信息的城市商圈区域范围划定
- 基于卷积神经网络的垃圾分类

- 基于图卷积网络的微博用户分群
- 基于微博大数据的疫情情感分析
- 基于自然语言处理的生物序列分析平台
- 人脸口罩在线检测系统
- 唐诗及诗人的文本挖掘与可视化
- 微博双旦娱乐偏好分析
- 新冠疫情相关的微博文本聚类分析
- 新闻联播讲稿分析
- 新闻语料数据抽取系统
- 行人及车辆检测系统
- 在线自然语言处理系统
- 自动写诗与鉴赏翻译系统

实验案例



基于短文本的情感分析综述

摘要：文本情感分析是近年来迅速兴起的一个研究课题。该文对文本情感分析的研究现状与进展进行总结。从情感分析的分类上可以分为基于情感辞典的方法以及基于机器学习的方法，重点阐述了围绕词典构建的一些方法和对于这两大类方法的介绍与分析，并介绍了文本情感分析的评测，提出了未来的研究方向。

关键词： 情感分析；情感辞典；机器学习

Research on Text Sentiment Analysis

Abstract: Text emotion analysis is a research topic that has arisen rapidly in recent years. This paper summarizes the present situation and progress of text emotion analysis. From the classification of emotion, analysis can be divided into based on the emotional dictionary method and machine learning-based approach.It is focusing on elaborating on the dictionary to build some of the methods and the introduction of these two methods and analysis. Besides,it describes the text of the evaluation of emotional analysis , proposing the future research direction.

Key words: Sentiment analysis; Emotional dictionary; Machine learning

1 引言

感谢您的耐心聆听！

张华平
Email: kevinzhang@bit.edu.cn
微博：@ICTCLAS张华平博士
实验室官网：
http://www.nlpir.org

大数据千人会