



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)

2020-10



- 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；
- 课代表：杜伦 高玉箫



大数据分析与应用2020



该二维码7天内(10月17日前)有效，重新进入将更新



➤ 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) 北理工十三五教材

➤ 天猫 清华大学出版社旗舰店47.84; 京东有售53.80。



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
- ✓ 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。



➤ 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。



➤ 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。

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

- 5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
- 每组上台报告(问题定义, 技术原理, 演示), 考核要点: 深入浅出、新、权威、团队配合; 需要超出《大数据智能分析》、老师和以前同学的报告。先报告的得分优先。

➤ 大数据智能应用项目(交付物: 代码, 说明文档, 演示PPT, 论文) 50分

- 5人一组, 可自由组合, 需标明分工; 提前2周提交项目, 选拔15组进入专家组终极评优答辩。
- 考核要点: 工作质量(有用、有趣、落地) 工作量;
- 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)**





如何拿90+?

- 比例控制在10%；名额控制在15人
- 结课前2周提交大作业，选取前15组进入终极PK；
- PK赛（最后一周）：每组15分钟演讲+演示答辩，评审组打分；
- 将署名并入选北理工十三五教材《大数据智能分析》第二版



基于NLPIR引擎的语义增强分析架构



语义增强分析平台

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)

- ✓ 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选题

- 平台层：1. 经典深度学习平台：TensorFlow/PyTorch平台；
- 算法模型：2. 经典机器学习算法；3. 经典深度学习方法；4-5. 终身学习方法；
- 数据导入：6. 网络爬虫采集；7. OCR；8. 语音识别；9. 多文档解析
- 语言分析：10. 语言预训练模型与应用（GPT3.0; BERT; ELMo; ）；11. 汉语分词；12. 句法分析；13. 情感分析；14. 文本校对；
- 信息抽取：15. 新词发现；16. 关键词提取；17. 专用命名实体抽取；18. 信息过滤与脱敏；19. 计算机图像实体检测
- 文本挖掘：20. 文本分类；21. 文本聚类；22. 话题发现；23. 相似检测与比对；
- 知识关联：24. 知识图谱构建与推理；25. 社交网络搜索与挖掘；26. 图文转换：看图说话与自动配图；27. 多模态分析；



	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

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	基于情感维度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

实验案例

1

提取
的半
统的
足要
们为
其中
整理
等信

摘要：
机
各项先
速，但
在机器
组成，
间，就
器翻译
文将简
着，对
训练数

面向

摘要：
分类的主
另一个是
过滤停用
关键词：

Literatu

Abstract:
sentimen
the main
are two
characte
statistica
word seg
future re
Keywords

0.引言

随着
评测网站等发表对产品的

电子邮
圾邮件的产
究一种十分
圾邮件识别
垃圾邮件分
(1)了
在主要的反
(2)了
实现垃圾邮

1 绪论

1.1 研究

电子邮
过信息电子
络的电子邮

基于朴素贝

摘要：近年

Answering Sy

答系统允许用

定领域问答系

RDQAS)针对特

其具有较好的

的发展进行了

对国内外近年

关键词：问答

A Survey on

Abstract: Rec

医疗
关医疗数
浪尖,而将
应用价值
生物
齐对于构
医学术语
的文本聚
将文本映
特征向量
之间的相
病节点为

一、

1.1 传

这里

断,治疗

学影像等医疗信息都是为其前往的医疗机构所拥有,患者、其他医院的医生都很难获得这些信息。同时,由于医疗机构相互之间缺少互联互通,其导致的最为直接的结果就是,患者更

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词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；情感词典；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

Answering System (OAS) has been widely studied and applied in various



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



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY