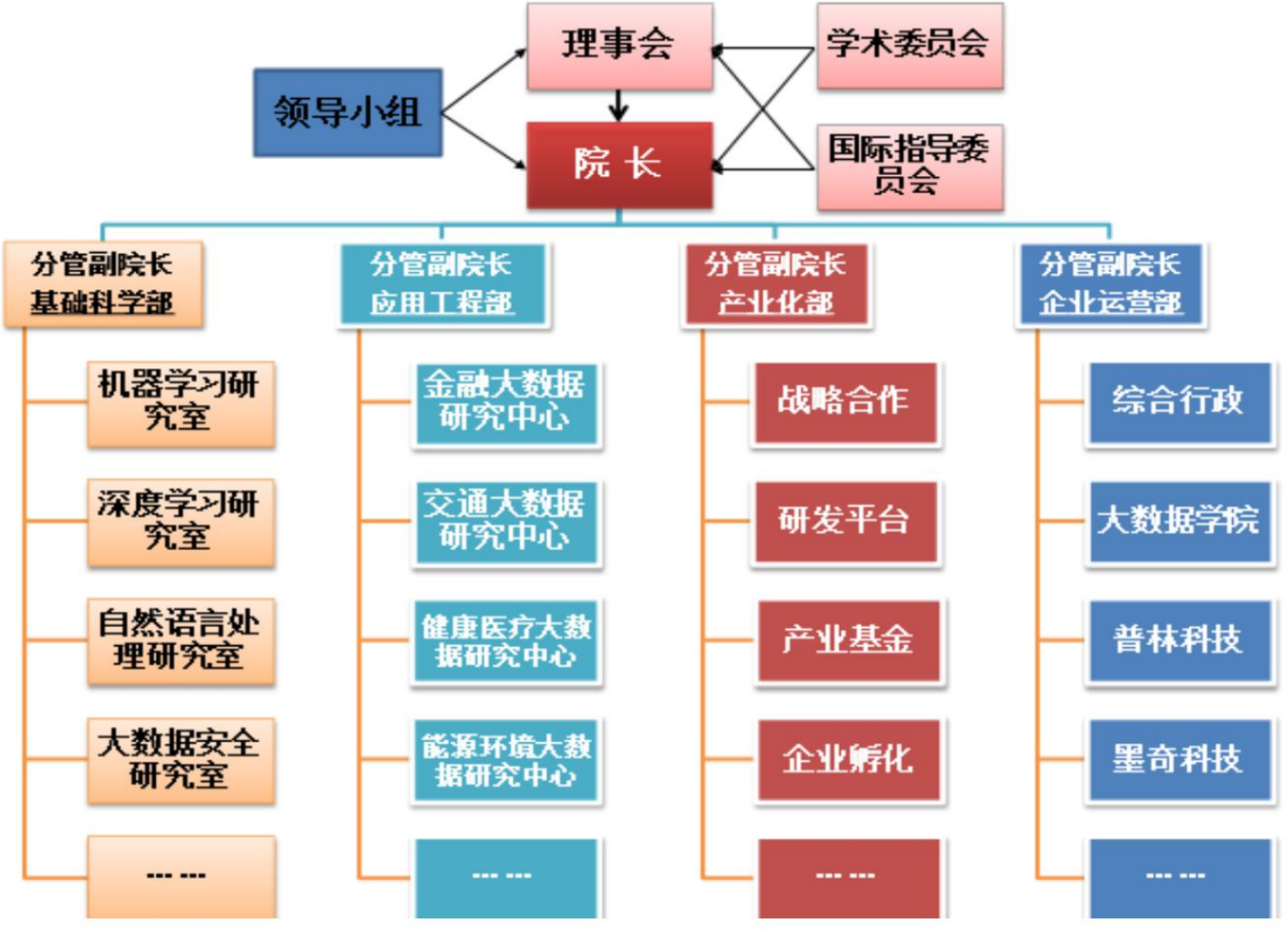
数据科学学科体系与实训平台建设

欧 高 炎
北京大数据研究院大数据教育研究中心主任
大数据教育联盟 秘书长
博雅大数据学院 院长
datascience@pku.edu.cn

北京大数据研究院发展历程

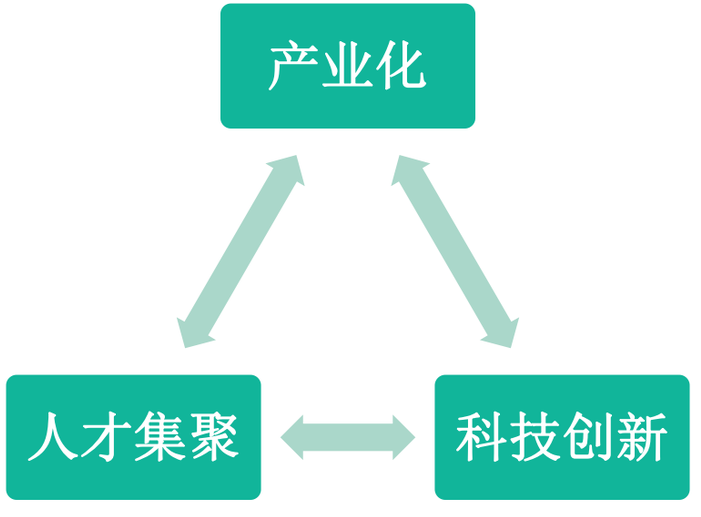
组织架构


领导小组
理事会
学术委员会
院 长
国际指导委员会
分管副院长 基础科学部
分管副院长 应用工程部
分管副院长 产业化部
分管副院长 企业运营部
机器学习研究室
深度学习研究室
自然语言处理研究室
大数据安全研究室
--- ---
金融大数据研究中心
交通大数据研究中心
健康医疗大数据研究中心
能源环境大数据研究中心
--- ---
战略合作
研发平台
产业基金
企业孵化
--- ---
综合行政
大数据学院
普林科技
墨奇科技
--- ---

BIBDR的发展目标

- 核心目标：支撑北京打造国际大数据产业化的引领示范区。
- 主要途径：通过体制机制创新，聚集和培养国内外大数据领域的一流人才，打造全国乃至全球的大数据科技创新中心。

大数据产业化的双轮驱动：
人才集聚与科技创新



组建一流的大数据研究团队

◆ **院长鄂维南院士，**千人计划，北京大学教授、元培学院院长，美国普林斯顿大学教授。973项目“非结构化数据分析”首席科学家。

◆ **高文院士，**大数据研究院学术委员会主任，国家自然科学基金委员会副主任，ACM/IEEE Fellow。

◆ **张平文院士，**大数据研究院学术委员会主任、北京大学学科建设办公室主任。

首批国际、国内人才集聚

- 国际人才引进
  - 郜骋　普林斯顿大学
  - 汤林鹏　普林斯顿大学
  - 章思鑫　纽约大学
  - 张立　爱荷华大学
  - 朱占星　爱丁堡大学
  - 李千骁　普林斯顿大学
  - 周亚俊　哈佛大学
  - ...
- 国内人才引进
  - 张志华　上海交通大学
  - 刘云淮　公安部三所
  - 王亦伦　电子科技大学
  - 严睿　百度
  - ...

在人才培养体系上率先突破

研究生

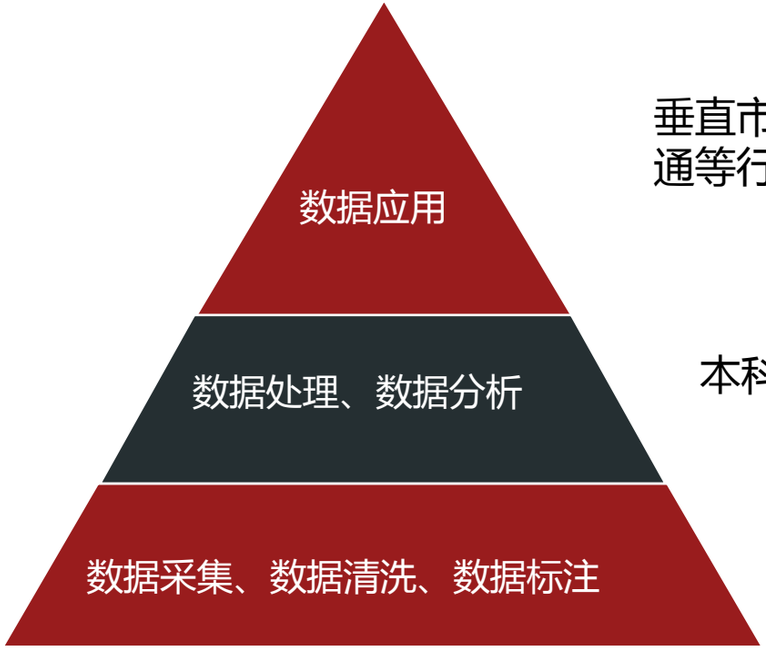
数据科学研究生专业建立，培养大数据方向的硕士、 博士

本科

数据科学本科专业，第一届学生2017年毕业。2016年成功申报全国首批“数据科学与大数据技术专业”

社会培训

与企业合作，开展相关培训2016年暑期学校、师资培训班、全流程实训平台等。成立博雅大数据学院、大数据教育联盟等。



大数据人才结构

垂直市场，培养金融、健康医疗、交通等行业大数据人才，众多行业企业

本科院校培养为主，1000+本科院校

职业院校培养为主，1000+所职业院校

提纲

1. 学科背景与专业介绍
2. 大数据师资队伍培养
3. 大数据课程产品建设
4. 数据嗨客：大数据教育实训平台
5. 大数据教育联盟

一、学科背景

- 2016年2月，教育部公布新增“数据科学与大数据技术”本科专业，二批次共**35所**院校获批。2017年**293**所院校申报。

- 2016年9月，教育部新增 “大数据技术与应用”专科专业，当前有**62所**院校开始设置该专业

“数据科学与大数据技术”(080910T)专业建设

培养目标：

本专业培养德、智、体、美全面发展，掌握数据科学的基础知识、理论、及技术。包括面向大数据应用的**数学、统计，计算机**等学科基础知识，数据建模、高效分析与处理， 统计学推断的基本理论、基本方法和基本技能。对自然科学和社会科学等**应用领域中大数据**的了解，具有较强的专业能力和良好外语运用能力，能胜任数据分析与挖掘算法研究和大数据系统开发的**研究型和技术型人才**

教学模式：

理论与实践结合，以数据为基础，实际问题为导向的**实践性**教学

“数据科学与大数据技术”培养怎样的人才？

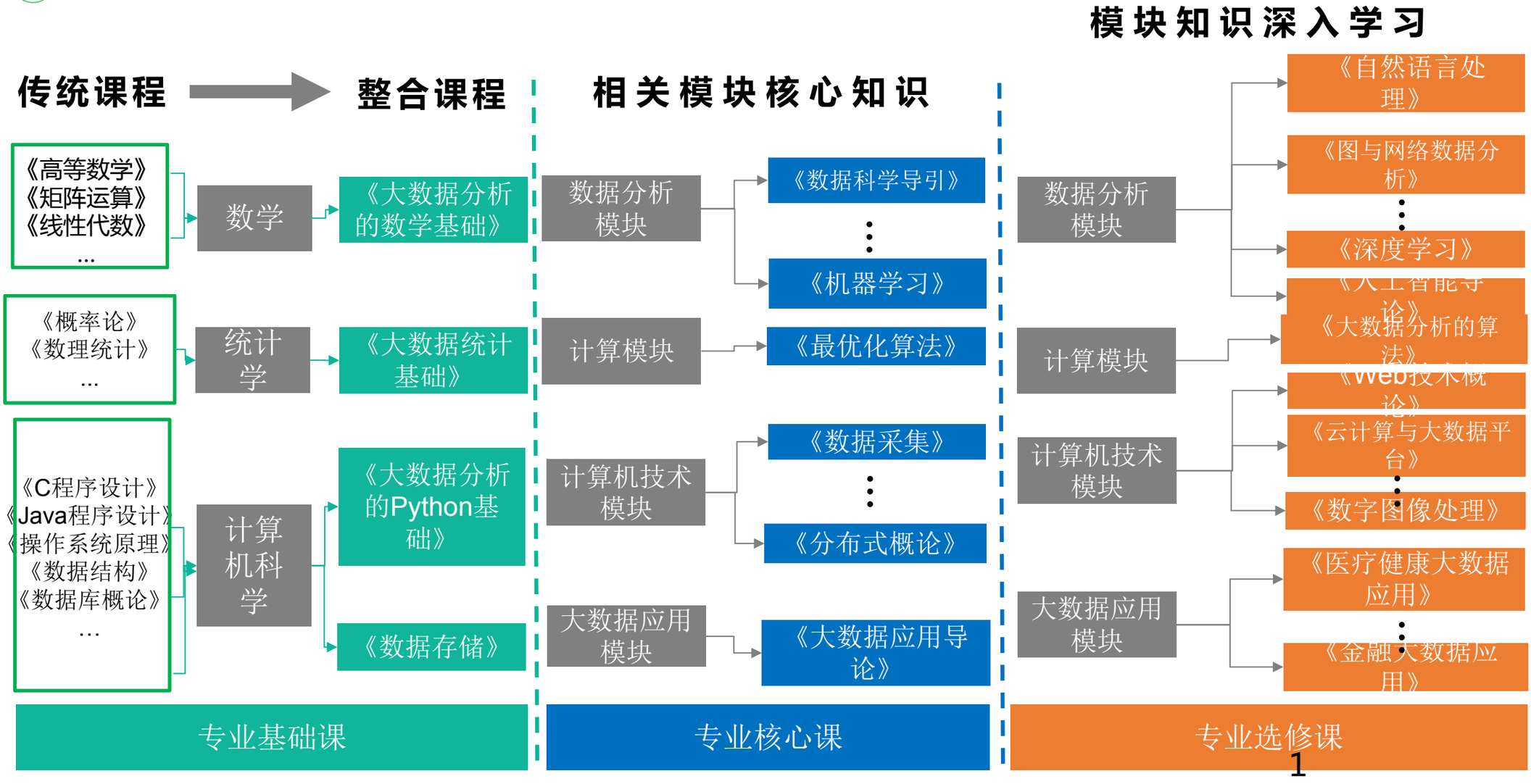
“数据科学学科强调培养具有多学科交叉能力的大数据人才。这样的人才应该具有以下三方面素质：一是**理论性**的，主要是对算法和模型理解和运用的能力；二是**实践性**的，主要是处理实际数据的能力；三是**应用性**的，主要是利用大数据的方法解决具体行业实际问题的能力。

培养这样的人才，需要数学、统计和计算机科学等学科之间的密切合作，同时也需要和产业界或其他拥有数据的部门之间的合作。

数据科学课程的开设，也需要采用新的模式，即理论课和实践课相结合的模式，就像物理、化学和生物课一样。这就需要提供相应的实验平台。这样的实验平台应该提供实际问题、实际数据和基本的处理工具。 “

------鄂维南院士
摘自《数据科学导引》序言

“数据科学与大数据技术”专业课程体系



“大数据技术与应用”(610215)专业建设

培养目标：

本专业培养掌握数据科学的**基础知识**及大数据相关技术，掌握大数据清洗和分析常用**工具**的使用，具有卓越的实践能力，能胜任**数据清洗、数据存储、数据分析与挖掘、大数据系统开发与构建**等工作的**专业应用型人才**。

教学模式：

以实践为主，以数据为基础，实际问题为导向、注重大数据处理和分析工具使用和实操应用的实践型教学

“大数据技术与应用”专业课程体系

语言和专业基础	数据采集、存储与处理	大数据分析、开发与应用
《大数据的语言基础Python》 《Linux系统基础》 《大数据的数理基础》 …	《大数据概论》 《人工智能概论》 《数据存储(MySQL)》 《数据采集与网络爬虫》 《数据清洗技术与工具》 《大数据处理工具(Excel/Weka/Pandas)》 …	《数据分析导论》 《大数据应用导论》 《数据可视化》 《Hadoop大数据平台基础》 …
专业基础课	专业核心课	专业选修课

二、师资队伍培养

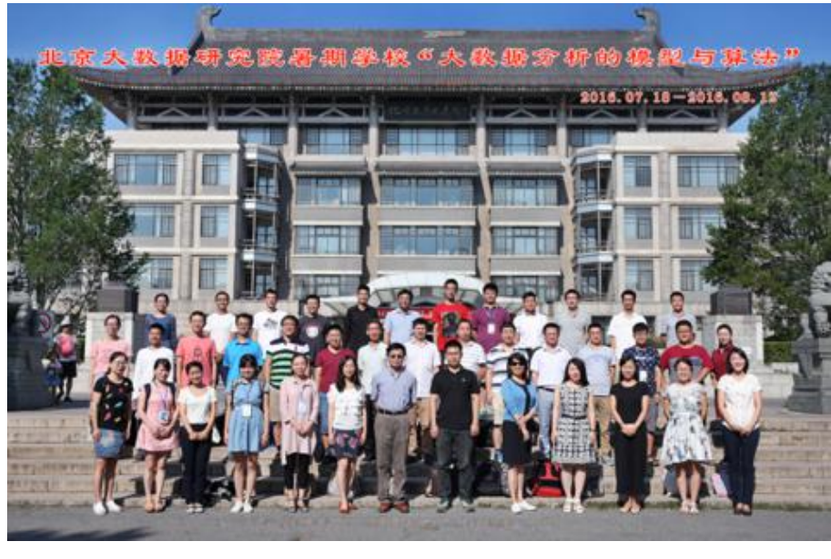
- 作为交叉型学科，大数据的相关课程涉及数学、统计和计算机等学科知识，对教师团队提出了更高的要求

- 北京大数据研究院和博雅大数据学院的师资培养工作
  - 2016年和2017年“大数据暑期学校”
  - 第1期和第2期数据科学与大数据技术专题培训
  - 多场企业定制培训
  - 已培养**500余名**高校师资和**数百名**企业数据分析人才

大数据暑期学校

- 2016大数据暑期学校：大数据分析的模型与算法
- 鄂维南（北京大数据研究院）沈佐伟（新加坡国立大学）纪　辉（新加坡国立大学）尹卧涛（加州大学洛杉矶分校）
- 张　潼（百度研究院）　时间：2016年7月18-8月12日

  报名500余人，录取70余名高校青年教师



2016暑期学校部分学员合影

- 2017大数据暑期学校：大数据分析的理论与应用

  时间：**2017年7月12-7月28日**

《机器学习的数学导引》　授课专家：鄂维南院士

《数据科学导引》　　授课专家：文再文教授　欧高炎

《知识图谱》　　授课专家：邹磊 教授

《大数据分析的语言与工具》 授课专家　北京大数据研究院专家

“大数据课程建设研讨会”　承办单位：南方科技大学、武汉大学、吉林大学等

交通大数据、金融大数据、健康医疗大数据等行业应用和前沿讲座

师资培训项目

- 北京大数据研究院专业师资团队授课
- 形式：每期招生30人以内，短期培训2-5天
- 结课：授予北京大数据研究院师资认证证书

2016年9月第1期
《数据科学导引》师资班

师资培训项目

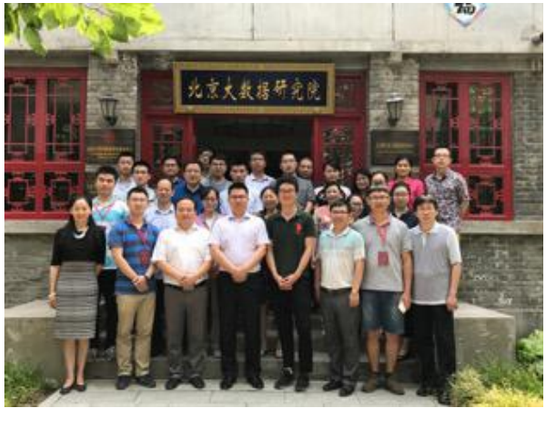
《数据科学导引》
2016年9月24日-9月28日

《大数据行业应用解析》
2017年4月10日-4月12日

《大数据分析的Python基础》
2017年4月14日-4月16日

《数据采集与网络爬虫》
2017年4月18日-4月21日

《数据清洗：技术与工具》
2017年5月12日-5月14日



《大数据分析的原理与技术》
2017年5月16日-5月19日

师资培训项目：活动预告

日期	培训课程	地点
10月21日-10月24日	大数据分析的模型与应用	北京
10月30-11月3日	大数据行业应用解析 大数据分析的Python基础	北京
11月5日-11月10日	数据清洗技术与工具 大数据分析的原理与技术	杭州
11月16日-11月19日	人工智能与深度学习	北京
11月27日-12月1日	数据采集与网络爬虫 数据清洗技术与工具	北京
12月18日-12月21日	大数据分析的原理与技术	北京

三、大数据课程产品

- 北京大数据研究院大数据系列课程产品
  - 《大数据分析的Python基础》《数据清洗》《数据科学导论》《数据采集与网络爬虫》
  - 研发中：《大数据应用导论》《金融征信》《大数据分析的数学导引》《数据可视化》等
  - **欧高炎、朱占星、董彬、鄂维南院士撰写的《数据科学导论》即将出版，高等教育出版社**
  - **《大数据分析的数学导引》教材北京大数据研究院与南方科技大学正共同撰写**
  - **欢迎更多联盟院校和企业加入，共同打造大数据学科精品教材**

- 成套的课程产品
- 讲义、实战案例、线上实训题库、课程教学视频

四、数据嗨客：大数据教育实训平台

官网：www.cookdata.cn 已开放免费注册

数据嗨客是北京大数据研究院和博雅大数据学院经过两年多研发的大数据教育实训平台。为高校大数据教学及企业数据人才培训提供线上实训环境及教学资料。让学生通过线上自主学习及实战演练，理解大数据科学的原理，掌握数据科学的知识体系，真实体验大数据建模分析的实际操作与演练过程。

目前，数据嗨客已支持北京大学、武汉大学、南方科技大学、北京信息科技大学等院校数千名师生开展大数据实践性教学

数据嗨客 HackData

- ✓ 权威的知识体系
- ✓ 在线大数据实训
- ✓ 随时随地线上练习
- ✓ 支持主流语言（Python和R）
- ✓ 自动效果评估
- ✓ 便捷的教学管理和在线作业管理

大数据教育实践探索

平台简介：

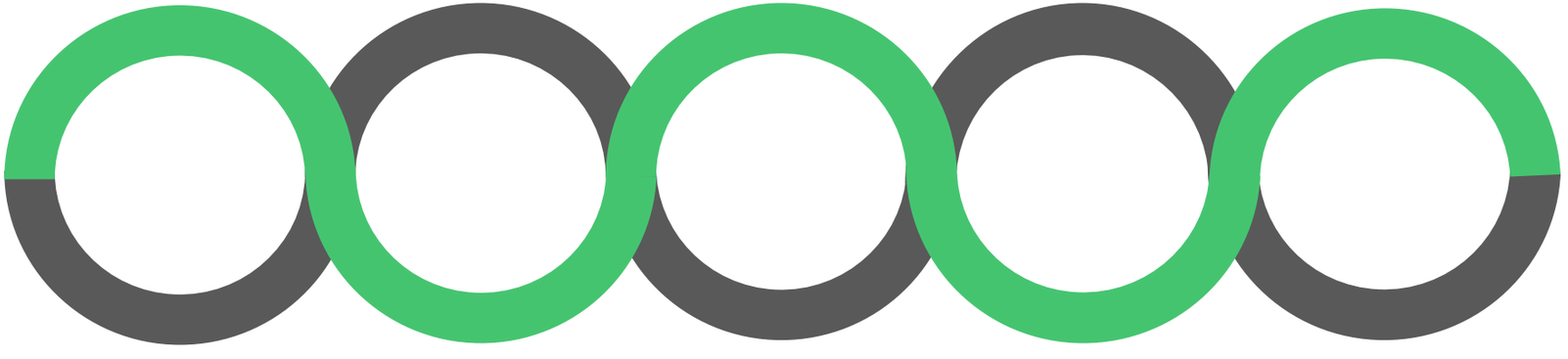
基于海量题库在线进行行业数据建模演练。涵盖多种体系，支持讨论和数据探索功能，实现自动模型训练、测试和评估，支持主流数据分析语言。

案例

专业的大数据教学管理和作业管理功能。除基本的教辅功能，支持针对大数据教学的在线作业发布、提交、评测和互动答疑等功能。

竞赛

利用用户画像技术构建大数据人才知识和技能图谱，实现人才和企业的有效匹配。



课程

全面的大数据知识讲义和丰富的案例资源。将大数据知识点进行有机整合，配套丰富行业案例，并提供大数据建模全流程指导。

教室

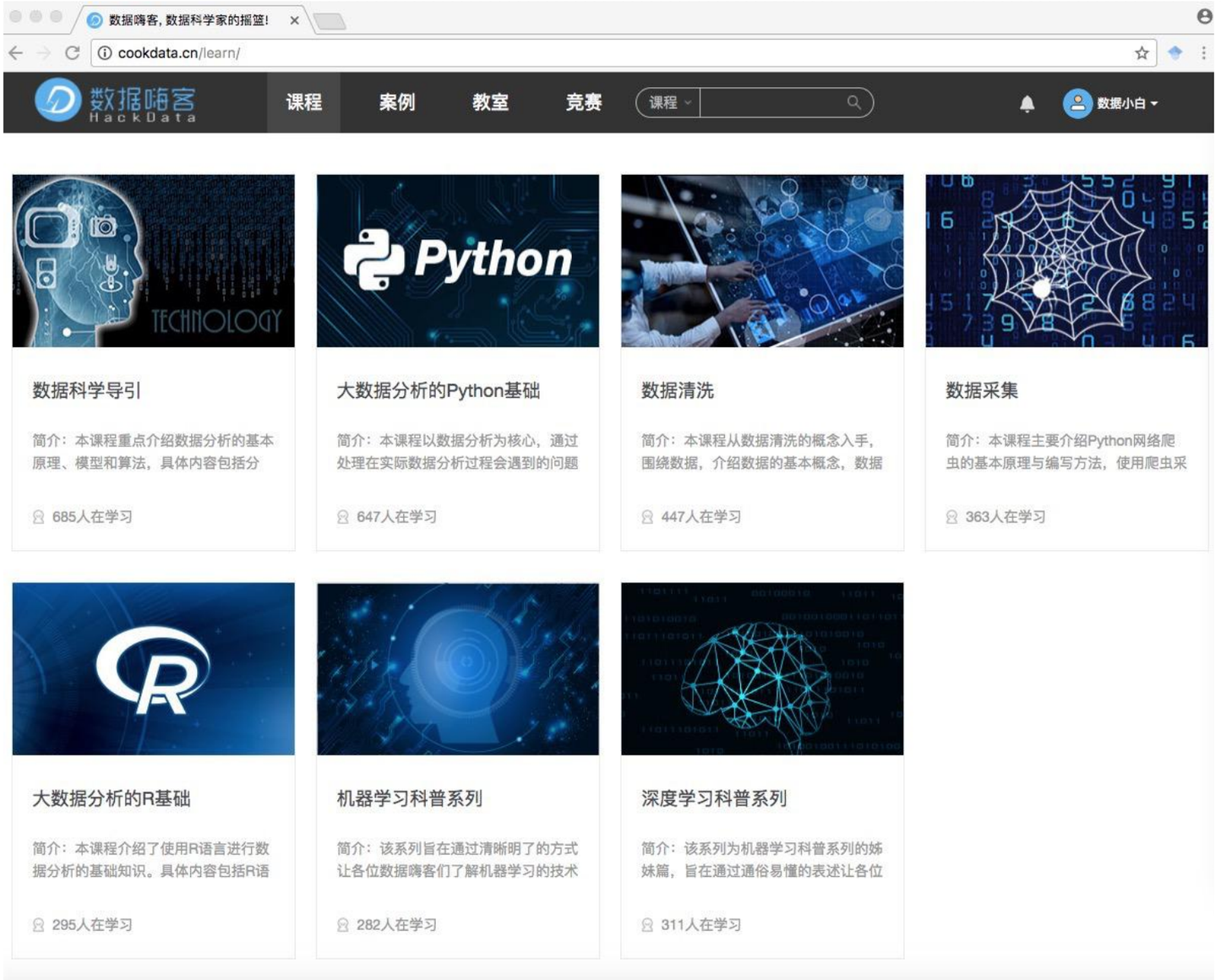
组织多种奖金丰厚的大数据竞赛，让大数据人才将所学知识和实战技能真正用于解决企业面临的挑战性问题。

工作

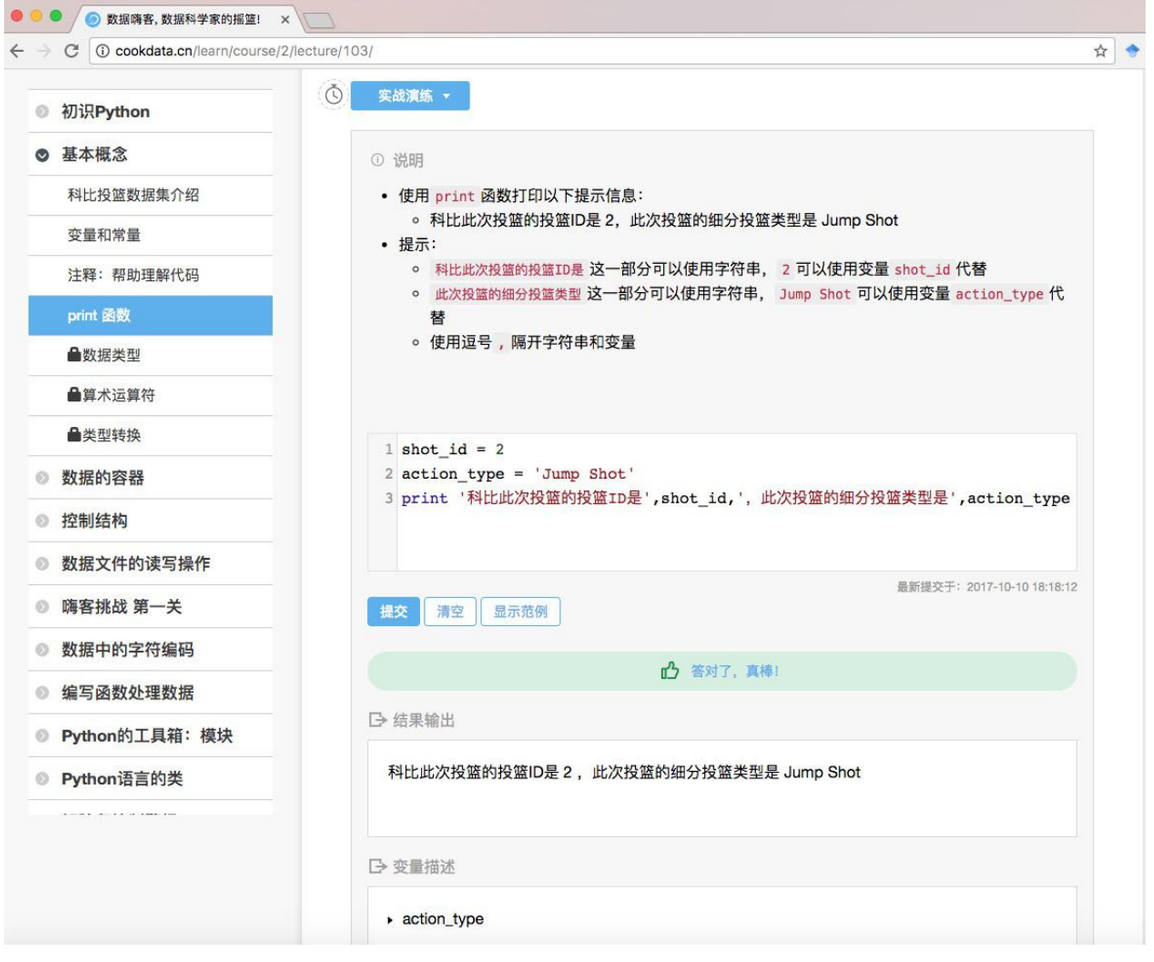
大数据教育实践探索

实操型课程

深入浅出的在线教学讲义和案例



随时随地在线学习并在线演练



独创的MagicFrame在线实训

- 理论和实操融为一体
- 在线测评学习效果
- 自定运行和记录学习过程

大数据教育实践探索

海量题库

提供大量在线数据实训题，进行大数据实战演练

#	题目列表	提交次数	通过率
430	流感患者检测	44	50.00%
420	钓鱼网站的识别	68	63.24%
414	南非西开普省冠心病分类	305	71.48%
401	酵母菌的蛋白质位置预测	89	95.51%
399	电离层反射的雷达波质量分类	57	96.49%
398	根据背景对人群的分类	57	98.25%
397	根据人名对徽章进行分类	43	100.00%
384	助教教学表现分类	40	97.50%
383	中文手机评论情感倾向判定	340	73.53%
377	Tictactoe游戏结果分类	55	98.18%
354	鸢尾属花的分类	63	88.89%
353	泰坦尼克号存亡者分类	45	93.33%
351	城市地表覆盖分类	34	100.00%
350	LED显示屏显示数字的辨别	24	95.83%
349	贷款违约预测	65	84.62%
346	红酒品质的分类	43	79.07%
328	人类皮肤检测	190	74.74%
322	大肠杆菌的蛋白质位置预测	300	76.00%

383. 中文手机评论情感倾向判定

时间限制:100s 内存限制:1024MB

题目描述

网购在我们的日常生活中起到了越来越重要的作用，当我
第一反应是参考相应的评论。每条评论中都有其情感倾向
中评。我们通过分析评论，有更大的概率做出较优的决策

本题目提供了一份从京东爬取的某款手机评论数据，每一
类别已被标注为好评(1)、差评(-1)或中评(0)。要求构建分
预测用户评论的情绪类别。

注意：用户评论为原始文本形式，需要用户自行完成中文

数据说明

数据包含Content和CLASS两列，分别为微博的文本内容

CLASS	Content
1	老板的发货速度确实挺快，东西也不
-1	谁买谁哭，用3天卡的跟什么一样，还不给退
0	马马虎虎

开始答题　　数据探索　历史提交　讨论

Python

```
class Solution(MLWorker):
    # 请在下面区域作答 #
    def train(self, dataframe_trainx, dataframe_trainy=None):


    def predictValue(self, model, dataframe_testx):
```

提交　清空　显示范例

大数据教育实践探索

测评报告

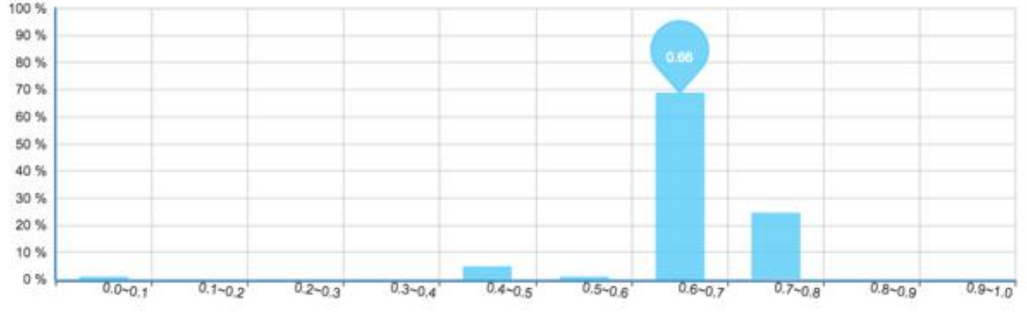
平台自动分析、评价数据分析与建模效果

模型评价

准确率 Accuracy	精确度 Precision	召回率 Recall	F1值 F1_score	对数损失 Log_loss	马修斯相关性系数 Matthews_corrcoef	ROC曲线下面积 ROC_AUC
0.66	0.63	0.66	0.64	--	--	--

分类 Classification	精确度 Precision	召回率 Recall	F1值 F1_score
-1	0.71	0.66	0.68
0	0.25	0.16	0.20
1	0.71	0.85	0.77

所有结果分布图 准确率



提交代码

```
class Solution(MLWorker):

    def train(self, dataframe_trainx, dataframe_trainy=None):

        from sklearn.tree import DecisionTreeClassifier
        from sklearn.feature_extraction.text import TfidfVectorizer
        import jieba

        x = dataframe_trainx[list(dataframe_trainx.columns.values)[0]].map(lambda xx:
        list_text = list(x)
        self.tfidf = TfidfVectorizer(max_df=0.95, stop_words=stopwords).fit(list_text
        array_trainx = self.tfidf.transform(list_text)
        array_trainy = dataframe_trainy.values.ravel()
        # 训练模型
        model = DecisionTreeClassifier().fit(array_trainx, array_trainy)
```

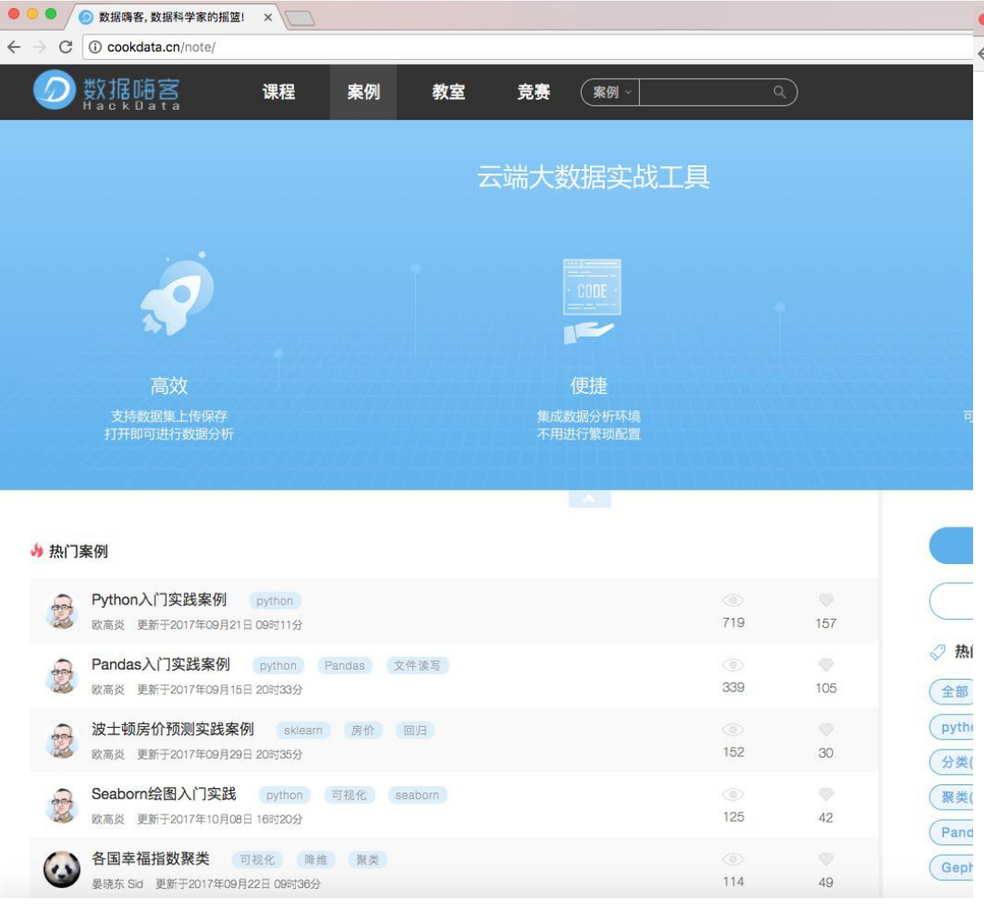
编译信息

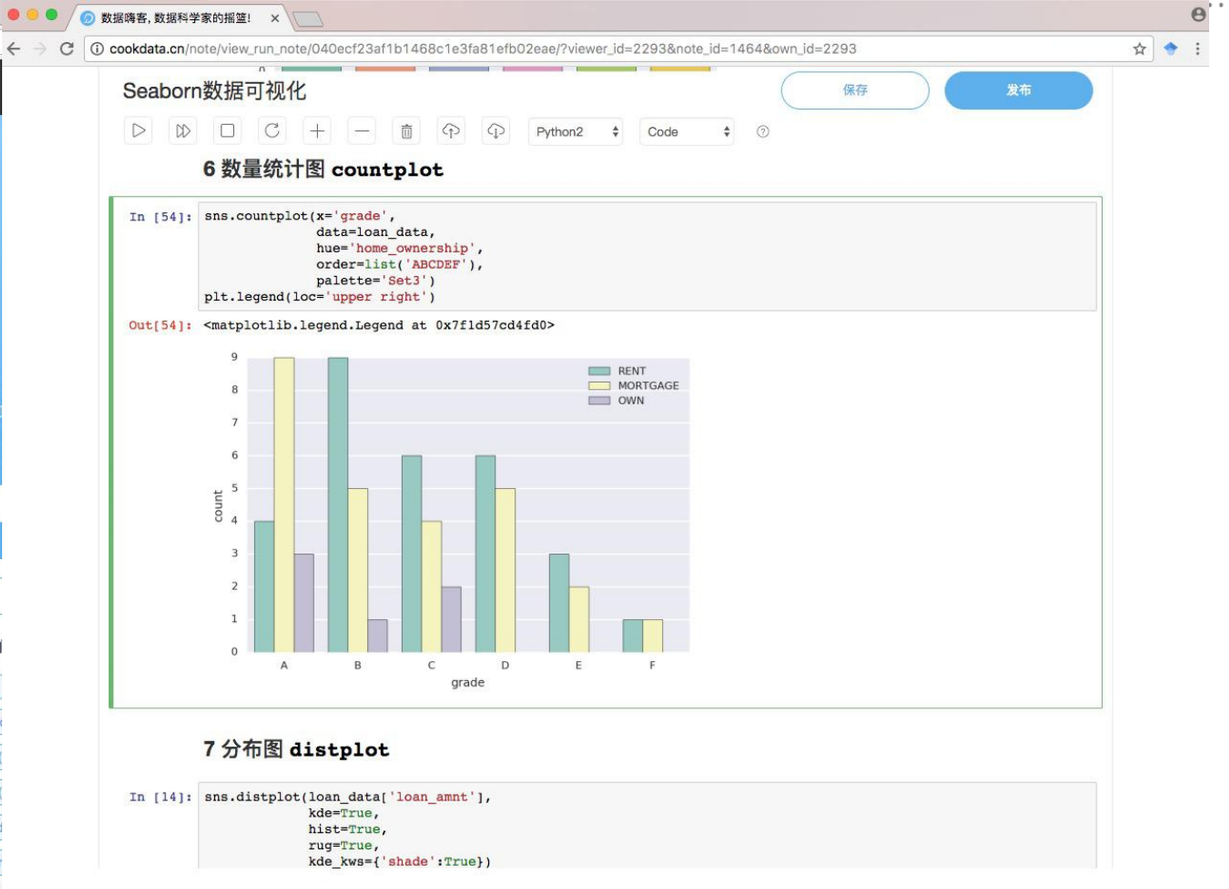
```
/tmp/p_sandbox/112300.py:13:80: E501 line too long (93 > 79 characters)
/tmp/p_sandbox/112300.py:19:1: E302 expected 2 blank lines, found 0
/tmp/p_sandbox/112300.py:26:70: E231 missing whitespace after ':'
/tmp/p_sandbox/112300.py:28:5: E301 expected 1 blank line, found 0
/tmp/p_sandbox/112300.py:37:1: E302 expected 2 blank lines, found 0
/tmp/p_sandbox/112300.py:50:80: E501 line too long (112 > 79 characters)
/tmp/p_sandbox/112300.py:52:80: E501 line too long (86 > 79 characters)
/tmp/p_sandbox/112300.py:64:80: E501 line too long (110 > 79 characters)
/tmp/p_sandbox/112300.py:72:5: E303 too many blank lines (2)
/tmp/p_sandbox/112300.py:73:80: E501 line too long (90 > 79 characters)
/tmp/p_sandbox/112300.py:75:80: E501 line too long (99 > 79 characters)
/tmp/p_sandbox/112300.py:91:4: W292 no newline at end of file
0.06     seconds elapsed
0        directories per second (0 total)
16       files per second (1 total)
```

大数据教育实践探索

案例

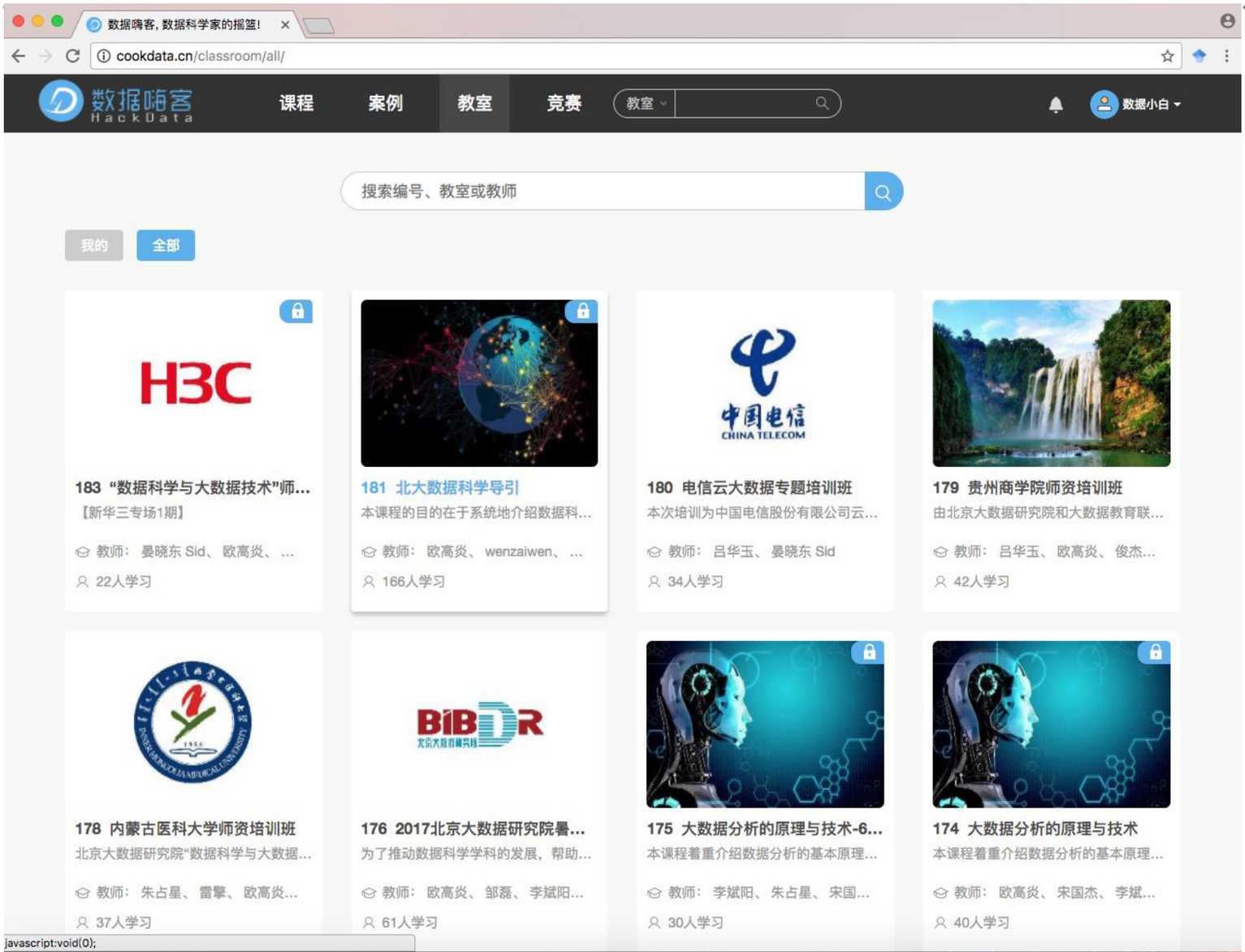
提供海量的大数据案例，数据处理和分析的云环境





大数据教育实践探索

线上教室　　组织管理在线教学



大数据教育实践探索

作业测评

自动生成班级作业测评报表，辅助教学

文本分析实战作业

截止时间：2016/12/12 23:00 编辑

最晚时间：2016/12/12 23:00

作业要求：本次作业给定一份中文手机评论数据，要求构建文本分类模型。评价指标为F1值。给分标准为：未通过 0分 F1 < 0.5 1分 0.5 =<F1 < 0.6 2 分 0.6 =<F1 < 0.7 3 分 0.7 =<F1 < 0.8 4 分 0.8 =<F1 5 分

作业完成情况

文本分析实战作业

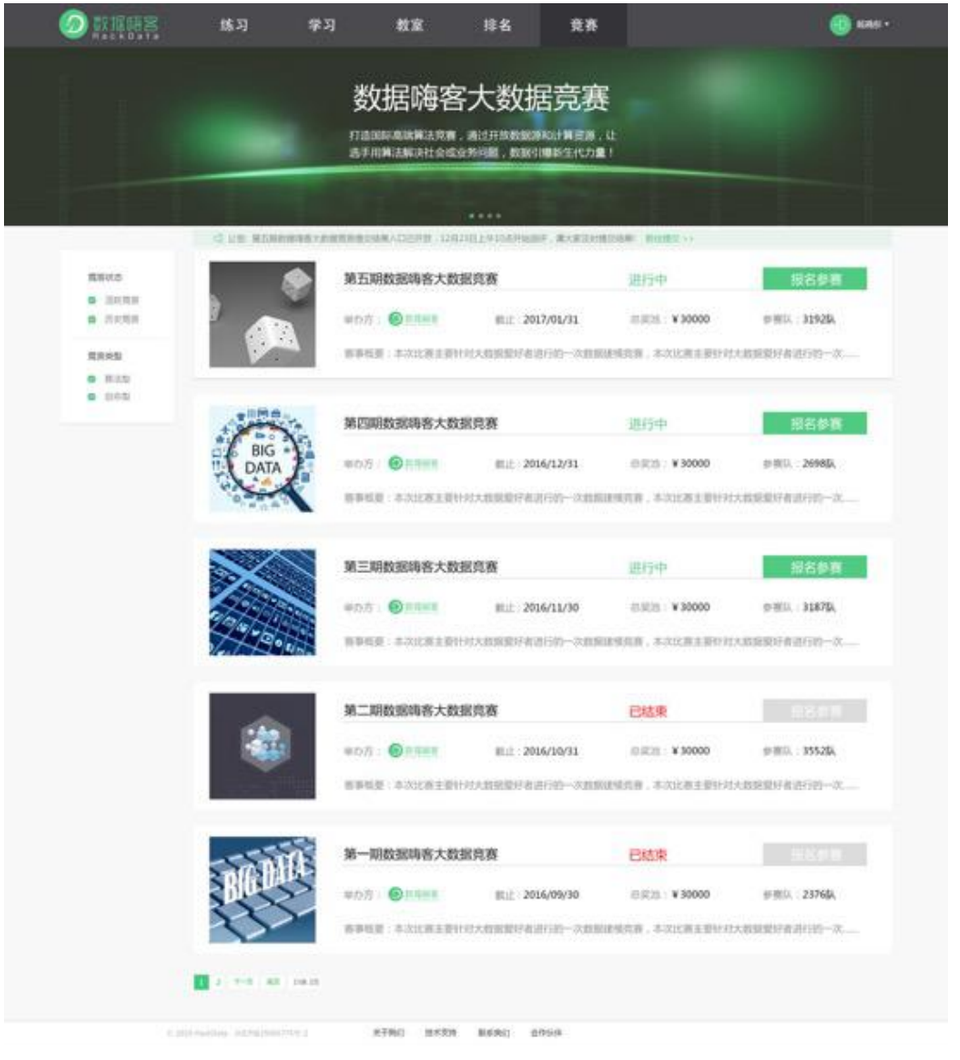
题目：中文手机评论情感倾向判定

用户	提交情况	提交次数	最后提交时间	状态	查看代码
wenzaiwen	未提交	0	-	成功0次	查看
jingweiw	未提交	0	-	成功0次	查看
龚佃选	未提交	0	-	成功0次	查看
王希舜	未提交	0	-	成功0次	查看
董彬	未提交	0	-	成功0次	查看
刘艳云	未提交	0	-	成功0次	查看
余冰	按时	39	2016-12-11 21:51	成功33次	查看
孟少帅	未提交	0	-	成功0次	查看
陈潇蒋	未提交	0	-	成功0次	查看
任惠霞	按时	11	2016-11-20 12:52	成功1次	查看
温见培	按时	5	2016-12-07 15:14	成功0次	查看
梁泽	按时	12	2016-12-09 01:58	成功4次	查看
赵星楠	按时	26	2016-12-10 17:18	成功10次	查看
陈龙	按时	25	2016-12-11 21:55	成功23次	查看

大数据教育实践探索

行业竞赛

帮助企业解决大数据的实际问题



大数据教育实践探索

能力鉴定　基于大数据的人才能力评价

排名	用户	做题数	提交次数	通过率
1	weiyunlei	204	2555	89.82%
2	贾晓东	130	717	52.16%
3	马艳艳	111	1164	68.27%
4	欧高贞	88	387	74.16%
5	杨嘉琪	72	178	82.58%
6	jiakang	82	205	80.98%
7	高佳佳	70	290	80.34%
8	李雪鹏	47	62	82.26%
9	张兆磊	47	146	87.67%
10	马俊	35	84	64.29%

1 2 3 4 5 6 下一页 尾页 257条 26页

- ✓ 讲义和案例的浏览记录
- ✓ 练习题目的效果排名
- ✓ 参与竞赛的排名情况
- ✓ 语言能力鉴定（Python和R等）
- ✓ 数据分析历史提交代码
- ✓ 题目和竞赛完成数量

大数据能力鉴定报告

五、大数据教育联盟

大数据教育联盟成立于2017年5月23日，由北京大数据研究院与北京大学元培学院发起并联合国内众多高等院校、科研机构和企事业单位共同组建，秉承实用、公平、开放和共享的基本原则，展开针对大数据专业人才认证标准研究、课程体系建设与推广、学术与人才交流等工作，切实推进大数据“产学研用”的无缝结合，促进商学互动，为国家大数据战略创造良好的生态体系，为大数据产业输送专业的人才。

BDEA 大数据教育联盟 Big Data Education Alliance

副理事长单位：北京大学、清华大学、中国人民大学、中山大学、复旦大学、贵州大学、对外经贸大学、中南大学、武汉大学、南方科技大学、微软加速器、京东金融、新华三集团

理事单位：北京邮电大学、北京信息科技大学、北京航空航天大学、澳门科技大学、电子科技大学、上海财经大学等超过百所院校和数十家企业

五、联盟公益计划

- 为联盟院校提供课程体系建设、课程开设咨询服务
- 与联盟高校共同举办大数据人才研讨会、就业对接会
- 与联盟高校共同申报、承接大数据项目
- 教材研发：与联盟院校联合出版具有影响力的大数据教材
- 向联盟高校教师免费提供大数据教育平台数据嗨客账号

北京大数据研究院公众号

数据嗨客 HackData

www.cookdata.cn

大数据教育联盟公众号


THANK YOU !