



数据分析及实践

Analysis and Practice of the Data

实验课

刘 淇

Email: qiliuql@ustc.edu.cn

课程主页:

<http://staff.ustc.edu.cn/~qiliuql/AD2021.html>



数据获取与管理实验

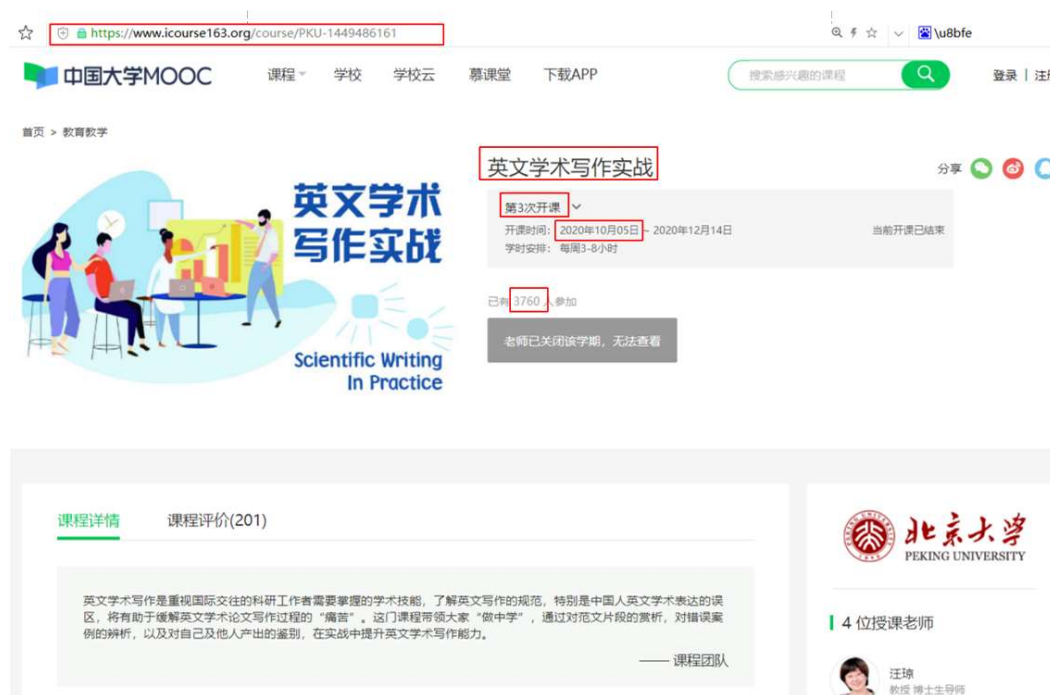
- 从以下两个实验任意选择一项完成
 - MOOC网站的课程信息爬取
 - 高考真题网站的信息爬取



实验二-MOOC

实验要求

- 给定网站: <https://www.icourse163.org>, 需要设计一个网站遍历策略, 爬取每门课程的相关信息, 记录于 json 文件中。部分信息标于红框中:





实验二-MOOC

☆ <https://www.icourse163.org/course/PKU-1449486161> 搜索 ☆ \u8bf8

中国大学MOOC 课程 学校 学校云 慕课堂 下载APP 搜索感兴趣的课程 登录 | 注册

首页 > 教育教学

英文学术写作实战 Scientific Writing In Practice

第3次开课
开课时间: 2020年10月05日 - 2020年12月14日
学时安排: 每周3-8小时
当前开课已结束

已有 3760 人参加
老师已关闭该学期, 无法查看

课程详情 课程评价(201)

北京大学 PEKING UNIVERSITY

□ 样例数据:

```
{
  "课程url": "https://www.icourse163.org/course/PKU-1449486161",
  "课程名称": "英文学术写作实战",
  "开课学校": "北京大学",
  "授课老师": ["汪琼", "范逸洲", "Torsten Juelich", "毛君"],
  "开课时间": "2020年10月05日",
  "开课次数": 3,
  "参加人数": 3760
}
```



实验二 -MOOC

注意事项

- 1. 一门课如果有多个开课学校，则取第一个开课学校
- 2. 一门课有如果多次开课记录，则取时间最晚的一次
- 3. 每位同学爬取至少500门课程的信息，课程种类不限
- 4. 保存到Jason文件的python代码，供参考（sample即为你解析得到的一个网页的数据字典）

```
import json

for url in urls:
    sample = get_obj(url)

    file = open('result.json', 'a', encoding='utf8')
    file.write(json.dumps(sample, ensure_ascii=False))
    file.write('\n')
    file.close()
```



实验二-MOOC

□ 提交要求

- 将爬虫代码和数据打包成一个压缩文件，发送给助教：
apdata2021@163.com
- 邮件标题: 姓名_学号_exp2_mooc
文件命名格式: 姓名_学号_exp2_mooc.zip
- 截止日期: 4月10日

□ 评分标准:

- 格式是否规范
- 提交是否及时
- 代码是否美观，能否运行



实验二-高考试题

实验要求

- 给定网站 <http://www.gaokao.com>，需要设计一个网站遍历策略，爬取高考真题试卷。

高考真题

按学科查找

语文真题

数学真题

英语真题

物理真题

化学真题

生物真题

地理真题

历史真题

政治真题

理综真题

文综真题

按年份查找

2020

2019

2018

2017

2016

2015

2014

2013

2012

2011

2010

2009

全国各地高考真题及参考答案

年级: 高一 科目: 语文 类型: 练习题 精确查找

试题导航

北京 | 上海 | 天津 | 重庆 | 广东 | 江苏 | 山东 | 浙江 | 湖北 | 四川 | 广西 | 湖南 | 辽宁 | 海南 | 宁夏 | 福建 | 甘肃 | 河北 | 江西 | 吉林 | 云南 | 河南 | 陕西 | 山西 | 安徽 | 新疆 | 西藏 | 贵州 | 青海 | 黑龙江 | 内蒙古 | 全国卷1 | 全国卷2 | 全国卷3

2020年	2019年	2018年	2017年	2016年	2015年	2014年
北京						
语文真题 答案	数学真题 答案	英语真题 答案	物理真题 答案			
化学真题 答案	生物真题 答案	地理真题 答案	历史真题 答案			
政治真题 答案						
上海						
语文真题 答案	数学真题 答案	英语真题 答案				
天津						
语文真题 答案	数学真题 答案	英语真题 答案	生物真题 答案			
物理真题 答案	化学真题 答案	地理真题 答案	历史真题 答案			
政治真题 答案						
重庆						
语文真题 答案	数学(理)真题 答案	数学(文)真题 答案	英语真题 答案			
文综真题 答案	理综真题 答案					
广东						
语文真题 答案	数学(理)真题 答案	数学(文)真题 答案	英语真题 答案			



实验二-高考试题

□ 该网站的试卷数据有两种保存格式： word和图片。

www.gaokao.com/e/20180614/5b22202144bba.shtml

按学科查找

语文真题

数学真题

英语真题

物理真题

化学真题

生物真题

地理真题

历史真题

政治真题

理科综合

文科综合

按年份查找

2020

2019

2018

2017

2016

2015

2014

2013

2012

2011

2010

2009

2018年浙江语文高考试题 (word版)

来源: 网络资源 2018-06-14 15:58:25

[标签: 浙江高考试题 2018语文高考试题]

2018年高考已经结束, 高考网整理了2018年各地高考试题、答案, 下面是2018年浙江语文高考试题(word版), 供大家参考。

2018年普通高等学校招生全国统一考试(浙江卷)

语 文

一、语言文字运用(共20分)

1. 下列各句中, 没有错别字且加点字的注音全都正确的一项是(3分)

A. 从懵(měng)懂的幼儿到朝气蓬勃的少年, 从两鬓斑白的青年到成熟的中年, 最后步入两鬓(bìng)斑白的老年, 有序变化是生命亘古不变的主题。

B. 虽然语言系统有自我净化能力, 随着时间的推移, 会分层过滤, 淘尽渣滓(zǎi), 浮(cù)炼真金, 但是当下网络语言带来的一些负面影响仍不容小觑。

C. 江上一个个漩涡, 似乎在仰首倾听清晨鸟鸣; 那些堆垛(duī), 战车, 均已废弛; 鸟鸣声穿过山风烟霭, 落满了山峦; 遍野麦浪, 渐成廖(liáo)廓之势。

D. 对于那些枉顾道德与法律铤而走险的电商平台, 有关部门必须给予相应的惩(chéng)罚, 否则难以制止种种薄(bào)客羊毛的恶劣行为。

阅读下面的文字, 完成2-3题。(5分)

在第55届博洛尼亚国际儿童书展上, 中国插画展现场观众络绎不绝, 显示出各界对中国插画现状与发展的关切。【甲】什么是插画? 插画就是出版物中的插图: 一本书如果以插画为主, 以文字为辅, 就被称为绘本, 顾名思义就是画出来的书。一本优秀的绘本, 可以让不识字的幼儿“读”出其中蕴涵的深意。【乙】在名画之下, 蝴蝶、花朵、叶子、大树等跃然纸上, 孩子可以识色彩、识物、识人、识事, 在学校里阅读的绘本, 父母在家里也可以和孩子一起阅读。如此一来, 孩子在幼儿园抑或在家里, 都拥有一个语言互动的环境。【丙】“绘本在儿童早期教育中的作用已被越来越多的人认识, 但绘本的发展还需加快步伐。”书

点击下载word版试题

2019年北京高考语文试题(图片版)(3)

来源: 北京高考资讯 2020-07-07 10:02:51

[标签: 高考语文试卷 高考试卷答案 2019高考试卷]

4. 下列对材料一和材料二的理解, 正确的一项是(3分)

A. 材料一分析了城市环境特点, 认为应该减缓城市化的步伐。

B. 材料二的引文表达了科学家对城市生物进化速度的忧虑。

C. 两则材料中关于城市化是人类文明的产物的看法是一致的。

D. 两则材料中关于热岛效应是否有利于生物生存的看法相似。

材料三

研究发现, 每个物种一旦成功适应城市生活的同时, 会有多个物种在当地消失; 而一个物种若过于迅速适应了城市生活, 也意味着众多个体要做出牺牲。城市化引发的生物快速进化往往要付出代价。

多伦多、波士顿等城市里的白车轴草, 为抵抗寒性而舍弃了释放氮化物的能力。释放氮化物可以抵御来自食草动物的威胁, 但寒性会降低。而在市中心, 城市高温使得积雪极易消融, 没有了积雪的覆盖, 植物就难以抵御严寒。一项新的研究表明, 包括徒步旅行在内的城市活动, 正在促使世界各地的哺乳动物在夜间变得更加活跃, 呈现出夜行性增强的趋势。夜行性增强会带来一系列的负面影响, 包括习性的改变、繁殖能力的降低等。关于纽约市各公园白足鼠的研究发现, 相比乡村白足鼠, 城市白足鼠体内涉及脂肪代谢的基因出现过度表达。此种进化选择极有可能会与在城市中能够更容易吃到人类丢弃的油脂、吃剩的比萨饼和芝士汉堡有关。自1940年以来, 意大利城市地区家猫的骨骼体积在不断增大。这为种群受路灯影响、路灯会吸引并聚集大量的大型昆虫, 随着世代更替, 吃虫为生的猫科动物具有优势。

城市中的生物进化与生物多样性密切相关。生物进化是一个难以操控、可预见性低的课题。加拿大多伦多大学助理教授马克·约翰逊说: “我们观察到, 一些物种在全世界大部分城市中都呈现出趋同进化。在那个城市, 物种未能顺利适应, 个中缘由目前还不得而知。”(取材于赵旭熙等的相关文章)

5. 根据材料三, 下列理解不符合文意的一项是(3分)

A. 白车轴草为抵御积雪的覆盖而舍弃了释放氮化物的能力, 这与城市高温有关。

B. 哺乳动物因夜行性增强而改变了习性, 繁殖能力降低, 这与人类的活动有关。

C. 城市白足鼠可能因为吃了比萨饼等食物, 涉及消化的某种基因出现过度表达。

D. 路灯吸引并聚集了大量的大型昆虫, 家猫因捕食它们而骨骼体积不断增大。

6. 就城市化与生物多样性的关系, 上面三则材料分别表达了什么观点? 说说这些观点对你认识这一关系有何启发。(7分)

语文(北京卷) 第3页(共10页)

北京高考资讯



实验二 - 高考试题

注意事项

- 1. MOOC项目与高考试题项目 **任选一个** 完成即可
- 2. 每位同学爬取 30份高考真题试卷，省份、科目、时间不限
- 3. 对于图片格式的试卷，
 图片的命名格式为：时间_地区_科目_计数.jpg (png)，
 如：2019_北京_语文_3.jpg
- 4. 对于Word格式的试卷，
 Word文件的命名格式为：时间_地区_科目.docx
 如：2018_安徽_语文.docx



实验二-高考试题

□ 提交要求

- 将爬虫代码和数据打包成一个压缩文件，发送给助教：
apdata2021@163.com
- 邮件标题: 姓名_学号_exp2_gaokao
文件命名格式: 姓名_学号_exp2_gaokao.zip
- 截止日期: 4月10日

□ 评分标准:

- 格式是否规范
- 提交是否及时
- 代码是否美观，能否运行

实验二-参考资料

- 实验二需要掌握 request库、正则表达式或 beautifulsoup库。
- 可以看相关博客入门，也可以阅读参考书籍：

