数据分析及实践

Analysis and Practice of the Data

实验课

刘 淇
Email: qiliuql@ustc.edu.cn

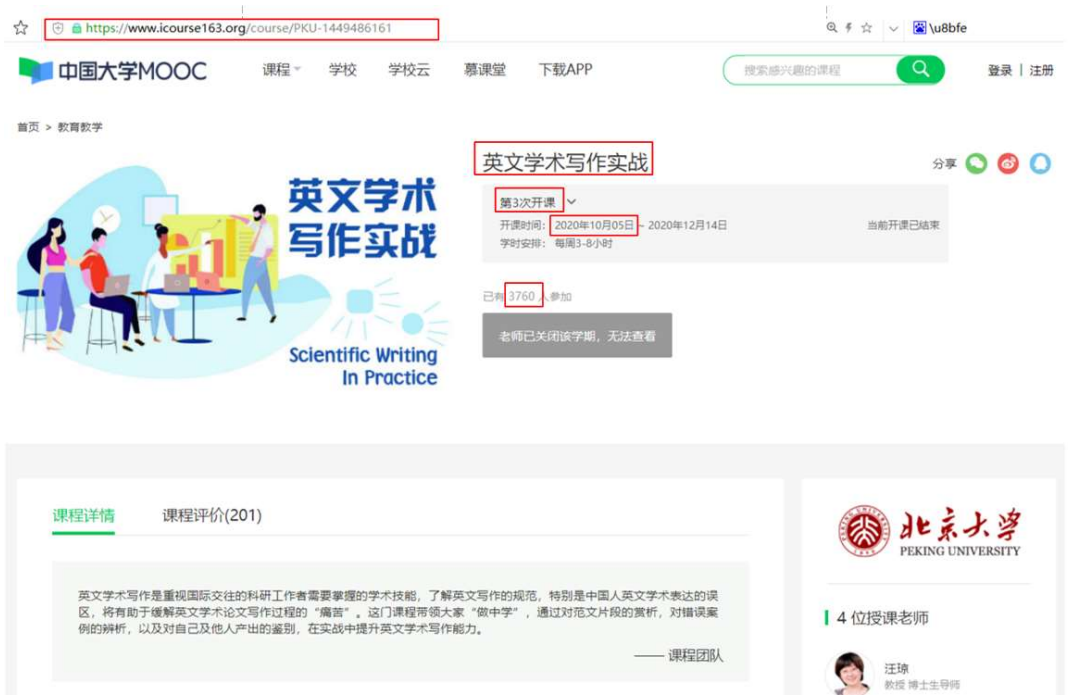
课程主页：
http://staff.ustc.edu.cn/~qiliuql/AD2021.html

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

- 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
}
```

json

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。

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
理综真题
文综真题

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（méng）懂的幼儿到朝气蓬勃的少年，从聘聘袅袅的青年到成熟的中年，最后步入两鬓（bing）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（zi），淬（cui）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. 江上一个个旋涡，似乎在仰首倾听清晨雁鸣；那些矮塔（die）、战车，均已废弛；鸟鸣声穿过山风烟霭，荡满了山峦；遍野麦浪，渐成燎（liáo）原之势。

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

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，中国插画展现场的观众络绎不绝，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，可以让不认字的孩子“读”出其中蕴涵的深意。【乙】在各色画笔下，蝴蝶、花朵、叶子、大树等跃然纸上，孩子可以对色彩、实物进行认知学习。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，孩子在幼儿园和成在家里，都拥有一个语言互通的环境。【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中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，每个物种每一次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地适应了城市生活，也意味着有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，为提高抗寒性而舍弃了释放氧化物的能力。释放氧化物可抵御来自食草动物的威胁，但抗寒性会降低。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，植物就难以抵御夜间冰冻。一项新的研究表明，包括徒步旅行在内的人类活动，正在促使世界各地的啮齿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酸消化的基因出现过度表达。此种进化选择极有可能与在城市中能够轻易吃到人类丢弃的油脂、吐司的比例增加有关。自 1940 年以来，意大利城市地区家蝠的颅骨体积在不断增大，这或许是受路灯影响。路灯会吸引并聚集大量的大型昆虫，随着世代更替，吸食力强的蝙蝠越发具有优势。

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破解、可预见性低的课题。加拿大多伦多大学动理教授马克·约翰逊强调说：“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。在部分城市，物种未能顺利适应，个中缘由目前还不得而知。”

（取材于赵熙熙等的相关文章）

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氧化物的能力，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，家蝠或因捕食它们而颅骨体积不断增大。

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）

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：