

法律声明

□ 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：大数据分析挖掘

■ 新浪微博：ChinaHadoop



第三讲



本地数据的采集与操作

--梁斌

目录

- 常用格式的本地数据读写
- SQL常用语法讲解
- Python的数据库基本操作
- 数据库多表连接用法详解
- 实战案例：欧洲职业足球数据库分析

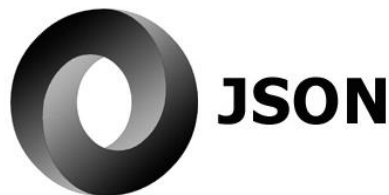
目录

- 常用格式的本地数据读写
- SQL常用语法讲解
- Python的数据库基本操作
- 数据库多表连接用法详解
- 实战案例：欧洲职业足球数据库分析

常用格式的本地数据读写

常用的数据分析文件格式

- txt
- csv
- json
- xml
- xls, xlsx
- HDF
- 其他可以转换成以上格式的数据文件
 - 如GIS中的.dbf可以导出成.csv文件



常用格式的本地数据读写

txt

示例代码： `01_txt_file_process.ipynb`

- 由字符串行组成，每行由EOL (End Of Line) 字符隔开， `'\n'`
- 打开文件
 - `file_obj = open(filename, access_mode)`
 - `access_mode: 'r' , 'w'`
- 读操作
 - `file_obj.read()` 读取整个文件内容
 - `file_obj.readline()` 逐行读取
 - `file_obj.readlines()` 返回列表，列表中的每个元素是行内容
- 写操作
 - `file_obj.write()` 将内容写入文件
 - `file_obj.writelines()` 将字符串列表内容逐行写入文件



常用格式的本地数据读写

txt (续)

示例代码：01_txt_file_process.ipynb

- 关闭文件
 - `file_obj.close()`



with 语句

- 包括了异常处理，自动调用文件关闭操作，推荐使用
- 适用于对资源进行访问的场合，确保无论适用过程中是否发生异常都会执行“清理”操作，如文件关闭、线程的自动获取与释放等
- `with open(filename) as f_obj:`
 - # 执行相关操作

常用格式的本地数据读写



.CSV

CSV (Comma-Separated Values)

- 以纯文本形式存储的表格数据（以逗号作为分隔符），通常第一行为列名
- 文件操作
 - numpy 的 `np.loadtxt()`，较复杂
 - 利用 **pandas** 处理，快捷方便
- 读操作
 - `df_obj = pd.read_csv()`，返回 `DataFrame` 类型的数据
- 写操作
 - `df_obj.to_csv()`

示例代码： `02_csv_file_process.ipynb`

常用格式的本地数据读写

Pandas

- 基于NumPy构建
- 索引在左，数值在右。索引是pandas自动创建的。
- 数据结构
 - Series，类似于一维数组的对象。
 - DataFrame，表格型数据结构，每列可以是不同的数据类型，可表示二维或更高维的数据



示例代码： 02_csv_file_process.ipynb

常用格式的本地数据读写

JSON (JavaScript Object Notation)

- 轻量级的数据交换格式
- 语法规则
 - 数据是键值对
 - 由逗号分隔
 - {}保存对象, 如{key1:val1, key2,:val2}
 - []保存数组, 如[val1, val2, ..., valn]

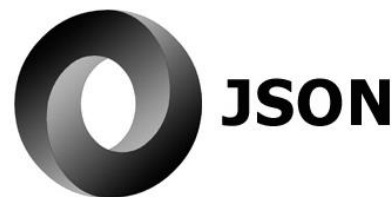


示例代码：03_json_file_process.ipynb

常用格式的本地数据读写

JSON (JavaScript Object Notation) (续)

- 读操作
 - `json.load(file_obj)`
 - 返回值是dict类型
- 类型转换 `json -> csv`
- 编码操作
 - `json.dumps()`
 - 编码注意
 - `ensure_ascii=False, encoding='utf-8'`



示例代码： `03_json_file_process.ipynb`

常用格式的本地数据读写

XLS/XLSX (Excel文件)

- 常用的电子表格数据
- 文件操作
 - 利用pandas处理，快捷方便
- 读操作
 - `df_obj = pd.read_excel()`，返回DataFrame类型的数据
- 写操作
 - `df_obj.to_excel()`
- 具体操作参考pandas如何处理CSV文件



常用格式的本地数据读写

HDF (Hierarchical Data Format)

- 存储不同类型的图像和数码数据的二进制文件格式
- 可以存储2种数据对象
 - 数组类型的数据
 - 目录结构的数据
- 安装h5py
 - `conda install h5py`
- 文件操作
 - `f_obj = h5py.File(file_name, mode)`
- 主要用于图像识别的数据存储
- 了解即可



目录

- 常用格式的本地数据读写
- **SQL常用语法讲解**
- Python的数据库基本操作
- 数据库多表连接用法详解
- 实战案例：欧洲职业足球数据库分析

SQL常用语法讲解

数据库的**CRUD**操作

- **Create**
 - `INSERT INTO table_name (column1,column2,column3,...)
VALUES (value1,value2,value3,...);`
- **Read**
 - `SELECT column_name,column_name FROM table_name;`
- **Update**
 - `UPDATE table_name SET column1=value1,column2=value2,...
WHERE some_column=some_value;`
- **Delete**
 - `DELETE FROM table_name WHERE some_column=some_value;`

目录

- 常用格式的本地数据读写
- SQL常用语法讲解
- Python的数据库基本操作
- 数据库多表连接用法详解
- 实战案例：欧洲职业足球数据库分析

Python的数据库基本操作

SQLite



- 关系型数据库管理系统
- 嵌入式数据库，适用于嵌入式设备
- SQLite不是C/S的数据库引擎
- 集成在用户程序中
- 实现了大多数SQL标准

示例代码： 04_sqlite_basic.ipynb

Python的数据库基本操作

SQLite



- 连接数据库
 - `conn = sqlite3.connect(db_name)`
 - 如果db_name存在，读取数据库
 - 如果db_name不存在，新建数据库
- 获取游标
 - `conn.cursor()`
 - 一段私有的SQL工作区，用于暂时存放受SQL语句影响的数据

示例代码： `04_sqlite_basic.ipynb`

Python的数据库基本操作

SQLite (续)



- CRUD操作
 - `cursor.execute(sql_str)`
 - `cursor.executemany(sql_str)` 批量操作
- `fetchone()`
- `fetchall()`
- `conn.commit()` , 提交操作
- 关闭连接
 - `conn.close()`

示例代码： `04_sqlite_basic.ipynb`

Python的数据库基本操作

其他常用数据库的连接

- Mysql
 - 主要面对互联网用户，比如建站等。
 - <https://dev.mysql.com/doc/connector-python/en/>
- PostgreSQL
 - Django推荐与PostgreSQL配合使用
 - Psycopg
 - <http://initd.org/psycopg/docs/>
- MongoDB
 - 分布式数据库
 - <https://docs.mongodb.com/getting-started/python/client/>

Python的数据库基本操作

其他常用数据库的连接

- Oracle
 - 适用于各类大、中、小、微机环境。它是一种高效率、可靠性好的 适应高吞吐量的数据库解决方案
 - <http://www.oracle.com/technetwork/articles/dsl/python-091105.html>

目录

- 常用格式的本地数据读写
- SQL常用语法讲解
- Python的数据库基本操作
- 数据库多表连接用法详解
- 实战案例：欧洲职业足球数据库分析

数据库多表连接用法详解

多表连接

- 查询记录时将多个表中的记录连接(join)并返回结果
- join方式
 - 交叉连接 (cross join)
 - 内连接 (inner join)
 - 外连接 (outer join)
- cross join
 - 生成两张表的笛卡尔积
 - 返回的记录数为两张表的记录数的乘积

示例代码：05_sqlite_join.ipynb

数据库多表连接用法详解

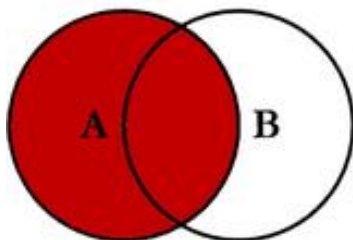
多表连接 (续)

- inner join
 - 生成两张表的交集
 - 返回的记录数为两张表的交集的记录数
- outer join
 - left join (A,B), 返回表A的所有记录, 另外表B中匹配的记录有值, 没有匹配的记录返回null
 - right join (A,B), 返回表B的所有记录, 另外表A中匹配的记录有值, 没有匹配的记录返回null
 - [注]目前在sqlite3中不支持, 可考虑交换A、B表操作

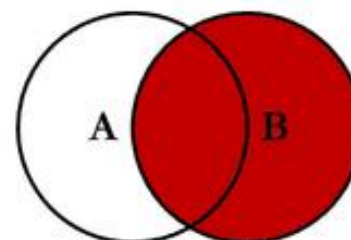
数据库多表连接用法详解

多表连接 (续)

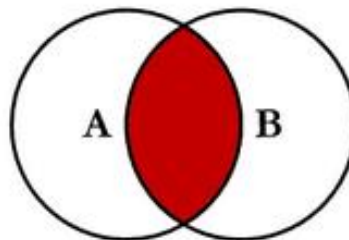
SQL JOINS



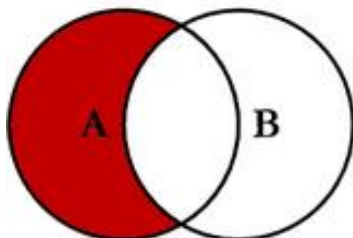
```
SELECT <select_list>  
FROM TableA A  
LEFT JOIN TableB B  
ON A.Key = B.Key
```



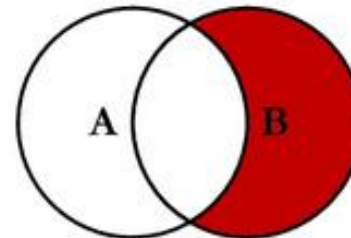
```
SELECT <select_list>  
FROM TableA A  
RIGHT JOIN TableB B  
ON A.Key = B.Key
```



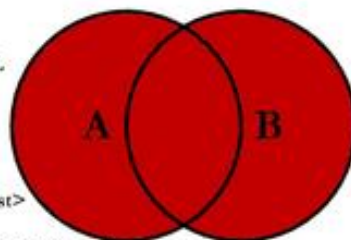
```
SELECT <select_list>  
FROM TableA A  
INNER JOIN TableB B  
ON A.Key = B.Key
```



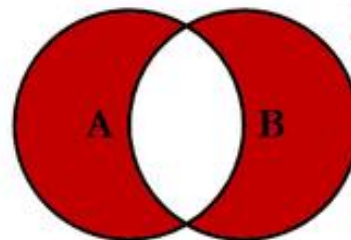
```
SELECT <select_list>  
FROM TableA A  
LEFT JOIN TableB B  
ON A.Key = B.Key  
WHERE B.Key IS NULL
```



```
SELECT <select_list>  
FROM TableA A  
RIGHT JOIN TableB B  
ON A.Key = B.Key  
WHERE A.Key IS NULL
```



```
SELECT <select_list>  
FROM TableA A  
FULL OUTER JOIN TableB B  
ON A.Key = B.Key
```



```
SELECT <select_list>  
FROM TableA A  
FULL OUTER JOIN TableB B  
ON A.Key = B.Key  
WHERE A.Key IS NULL  
OR B.Key IS NULL
```

© C.L. Moffatt, 2008

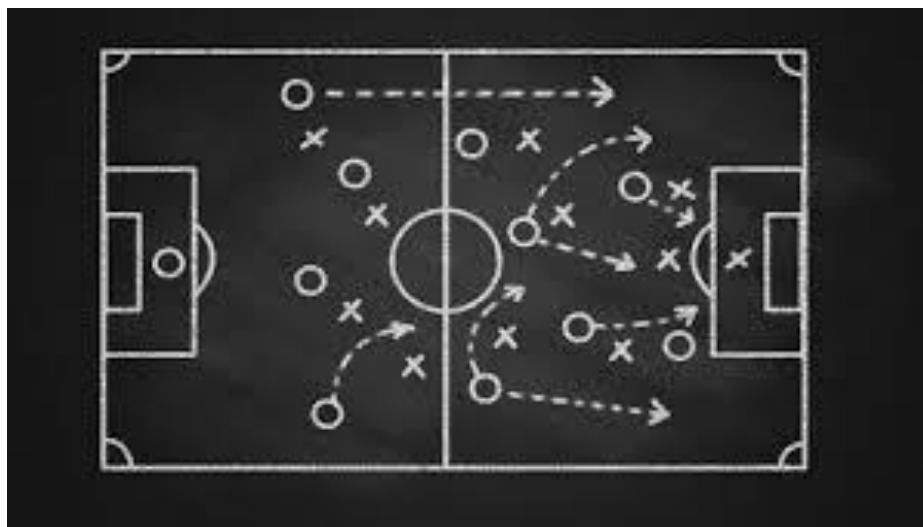
目录

- 常用格式的本地数据读写
- SQL常用语法讲解
- Python的数据库基本操作
- 数据库多表连接用法详解
- 实战案例：欧洲职业足球数据库分析

实战案例

项目介绍

- 项目地址：<https://www.kaggle.com/hugomathien/soccer>
- 欧洲足球数据库分析 (European Soccer Database)



示例代码：lecture03_proj.zip

实战案例

项目介绍


- 25,000+ 球队数据
- 10,000+ 球员数据
- 11个欧冠国家
- 2008-2016赛季
- 球员和球队属性数据来自EA FIFA电子游戏，包括每周的更新数据
- ...









示例代码：lecture03_proj.zip

实战案例

项目介绍

- 数据库结构

 database.sqlite

-  Country
-  League
-  Match
-  Player
-  Player_Attributes
-  sqlite_sequence
-  Team
-  Team_Attributes

- 数据表结构

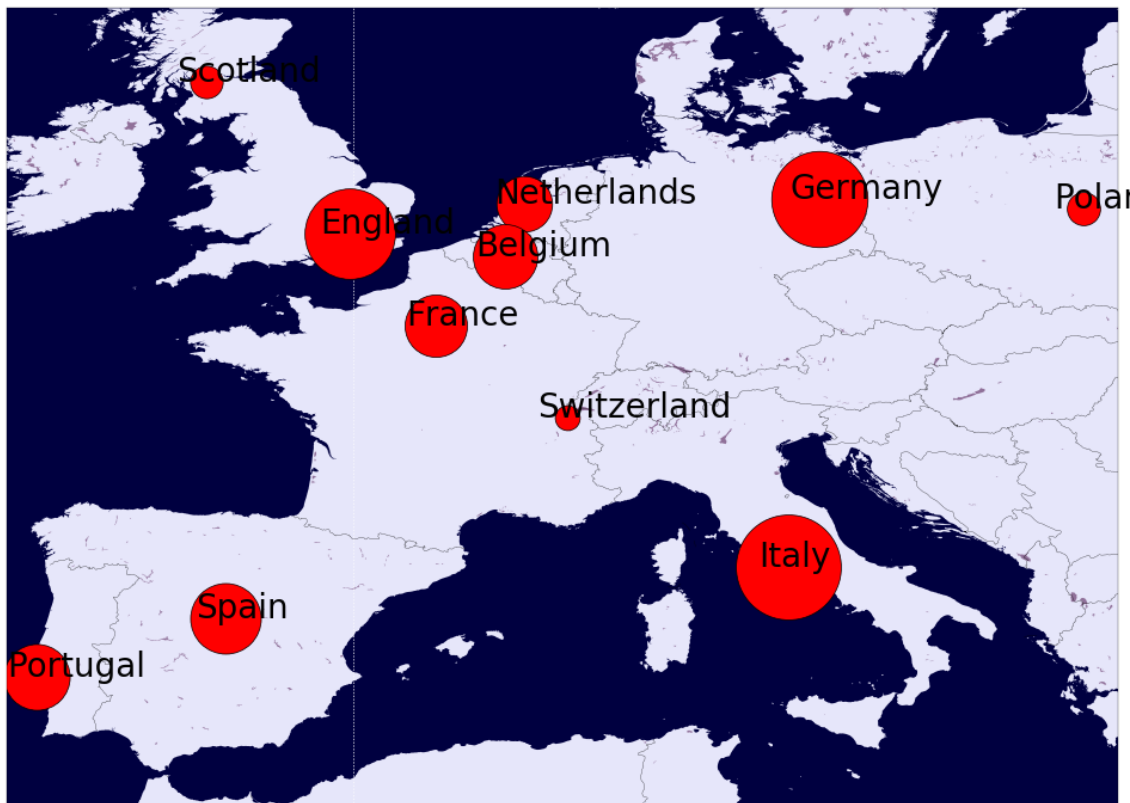
table_structure.csv

示例代码：lecture03_proj.zip

实战案例

项目目的

- 掌握Python的数据库连接
- 分析数据库信息
- 地图数据可视化



准备工作

- `conda install -c anaconda basemap=1.0.7`

示例代码：lecture03_proj.zip

实战案例

涉及知识点

- 模块化项目
- Python的SQLite连接
- 列表推导式、字典推导式
- DataFrame数据结构
- CSV数据生成
- Matplotlib简单的地图可视化

示例代码：lecture03_proj.zip

参考

- Python的文件读写

<https://docs.python.org/2/tutorial/inputoutput.html>

- Pandas的IO工具

<http://pandas.pydata.org/pandas-docs/stable/io.html>

- SQLite中的多表连接

https://www.tutorialspoint.com/sqlite/sqlite_using_joins.htm

- sqlite3模块

<https://docs.python.org/2/library/sqlite3.html>

- Python的中文编码

<http://blog.csdn.net/liuxincumt/article/details/8183391>

参考

- 项目代码参考自

<https://www.kaggle.com/doctorclo/d/hugomathien/soccer/can-you-be-a-good-football-player/notebook>

疑问

□ 问题答疑：<http://www.xxwenda.com/>

■ 可邀请老师或者其他回答问题

小象问答 @Robin_TY

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop

