Python设计方案

张晨峰

2015年9月3日

1 设计目标

- 数据导入和导出功能
  - 数据导入，包括从不同格式的文件及数据库导入数据
  - 数据导出，包括导出数据到不同格式的文件及数据库
- 数据预处理、清理和结构转换
- 探索性数据分析
- 数据挖掘
- 数据模型

2 设计方案

2.1 数据导入和导出

数据导入的主要任务是从不同格式的数据文件和数据库导入数据。

- 从Excel文件导入和导出数据
  - 构建Excel类
  - 常用方法包括连接Excel文件，从Excel文件导入数据及导出数据到Excel文件
- 从MongoDB数据库导入和导出数据
  - 构建MongoDB类
  - 常用方法包括连接MongoDB数据库，插入数据、查询数据和更新数据
  - 查询数据导出到DataSet类
- 爬虫程序

2.2 数据预处理

数据预处理的主要任务是从原始数据中提炼数据，并且储存为标准数据格式。

- 数据表单的根类DataSheet
  - 扩展不同的子类，应对不同的输入文档形式
  - 根类的属性是原始导入的数据
  - 数据表单类最终转换为标准数据格式类StandardFormatData
- 标准数据格式类StandardFormatData
  - 属性包括有序字典（储存数据）
  - 常用方法包括标准数据格式到数据集类DataSet、数据库MongoDB格式类MongoDBFormatData的转换
- 数据集类DataSet
  - 属性包括数据表
  - 常用方法包括数据集的信息提取和分析
- 数据库MongoDB格式导入根类MongoDBFormatImportData
  - 扩展不同的子类，应对不同的集合
  - 统一的输入插入格式

2.3 数据清理和结构转换

- 主要应用在DataSet中，利用NumPy、SciPy及Pandas包进行数据清理和结构调整
- 主要是进行Series和Dataframe的变换

2.4 探索性数据分析

2.5 数据挖掘

2.6 数据模型