

Python在大数据分析中的应用研究

李天倚

武昌职业学院 湖北武汉 430202

摘要:以Python为基础的开发工具,对于提高大数据的分析和处理等具有重要意义。将Python的优势与其自身优势相融合,从“快速”的角度出发,对其进行优化和改进,并通过可视化和数据分析,提升其在实践中的使用效率。在数据编程之下,Python语言可以对信息获取、信息存储等方面进行数据处理,以大数据为依据,可以提升数据处理和分析的能力。

关键词:大数据分析; Python语言; 应用分析

引言:

伴随着信息技术的不断发展和科学技术的不断提高,21世纪将会成为一个海量的信息的载体。在当今“大数据”的社会环境中,每一个学科的发展都必须依靠数据的解析来实现,而不再是单纯地依靠“经验”或“直觉”来进行决策。数据处理是大数据研究中不可或缺的一部分。资料处理就是运用恰当的统计学的手段,将所搜集到的资料加以整理,然后提取出有价值的资讯,进而得出结论。数据分析的工作就是从这些庞大且杂乱无章的数据中,寻找出那些有价值的数据,为这些数据赋予新的含义,并为决策者们提供一些建议。

一、Python语言及特点

Python是荷兰数学家吉多·凡·罗斯姆(GuidoVanRossum)发明的一种被广泛使用的跨平台的高级编程语言。Python的设计理念是,它的设计理念是:代码易读,并且使用简单明了的语言,Python从1991正式发行到现在,已经经历了几次修改,添加了很多新功能,并且越来越简单,越来越标准,现在已经在系统管理的工作和Web编程中得到了广泛的使用。Python是一个杀手锏,用来进行数据分析,在2017年,Python以一种令人眼花缭乱的方式成为最流行的程序设计语言,深受程序员们的喜爱。Python拥有大量的库,在进行快速开发时,表现出了极大的优势,在数据分析、数据科学、人工智能等学习工作中,Python是仅次于C++和Java的第三大语言^[1]。

Python语言的特性如下:

作者简介:李天倚(1989.10-),女,汉,湖北襄阳,硕士研究生,讲师,研究方向:信息技术与课程整合,计算机应用。

1.易于学习与其他的程序设计语言(例如C)相比,Python的程序设计很简单,很好地入门,很适用于新手。举个例子,我们想要实现一个特定的功能,Python所需要的代码只有C的十分之一,而用Python编写的话,可以极大地提升工作效率,这也是Python最受欢迎的地方。

2.在PythonObject中,所有东西都是Object。程序导向和物件导向都是支援的。一个程序建立在一个对象之上,这些对象包含了大量的数据和函数,Python相对于C++、Java等其他主流语言来说,具有超乎寻常的功能,同时也非常容易实现OOP。

3.通用性强,只要有对应的Python解释工具,Python软件就能在多种平台上正常工作,如Linux, Windows, Android等。Python的语法简洁明了,使用起来非常容易上手。它的语法规则清晰,给人一种自然语言的感觉,降低了学习难度。Python可以运行在多个操作系统上,包括Windows、Linux、Mac等。无论是在个人电脑上还是在服务器上,都可以轻松地使用Python进行开发。

4.Python标准类库非常的强大,能够为您提供大量的信息,比如regex表达式,数据库,web浏览器,XML,XML-RPC,HT.ML,WAV文件,密码系统,GUI,以及其他系统相关的功能。

5.Python标准的编码使用了强迫的缩排法,提高了编码的易读性。

二、Python语言的优势分析

首先,Python语法简洁清晰,代码易于阅读和理解。相比其他编程语言,Python的代码行数更少,具有更高的可读性和可维护性,有助于提高开发效率。其次,Python拥有丰富的第三方库和模块,可以方便地实现各种功能,如数据分析、机器学习、网络编程等。这些库和模块的存在极大地减少了开发人员的工作量,可以快

速实现复杂的功能。此外, Python语言可以在多个操作系统上运行, 包括Windows、Linux、Mac等。这使得开发人员可以方便地在不同平台上开发和运行Python程序, 提高了程序的可移植性。最后, Python语言能够创建一个标准数据库, 着重于对数据库、表达式等方面进行全面的控制, 并且使用了强制缩进的方法, 增强了代码的可读性, 在计算机编程中, 常用的语言主要有Java、C、Python等, 语言的种类也很多, Python语言是一种较为简单的语言, 并且便于后期的程序维护和管理, 它的使用也较为普遍。

三、大数据分析中Python语言的应用

1. 建立文件

它是一个利用软件工具来获取网页资料的程式。在设计网络爬取器时, 可以从信息页中获取信息数据, 使用lxml对有关的数据信息进行过滤, 将其存储到电脑硬盘。由于Python是一种面向对象的编程工具, 因此, 它被广泛地用于混合料的自动设计。由于程序设计技术日益多样化, 因此Python的类库变得更加强大。Python既可以独立使用, 也可以与Django结合使用, Python自身也有一些独特之处, 对于具体的使用, 我们可以使用Python来使用缩进式来实现if语句。Python在实践中的运用, 可以使数据的编译器更加完美, 使数据的运行更加顺畅。创建了一个文档, 然后根据文档中的一些重要参数, 对这些参数进行了相应的操作, 并将这些参数存储到相应的存储空间中。

2. 大数据信息抓取与控制

在Python语言的开发和信息的处理中, 可以利用搜索引擎来提供一个地址, 比如在软件开发结束之后, 就可以利用百度来搜索前面的资料, 并创建一个连接渠道, 根据资料分析和资料爬取, 就可以对资料的规则和资料的处理等进行控制, 来实现资料的抓取, 在新设站点之后, 可以通过与其他站点进行协作, 使用网页爬行工具进行搜索, 增加资料抓取法则, 以完成资料的分析和资讯加工。

3. 爬虫信息获取

在Python语言的运用之下, 为了得到页面的资讯资料, 可以使用网络爬程式, 搭建搜寻引擎, 并且对URL资料进行剖析, 达成资料取得和资讯分析的目标, 获取了这些信息后, 就可以和目的信息进行对比和分析, 查找到相应的URL信息。在获取URL数据之后, 可以将数据存储在本地硬盘上, 并集成数据信息, 提升数据信息的分析能力。

4. 存储信息的方式

在大数据分析中, Python语言可以用多种方式来存储信息, 其中一些常用的方式包括: 1. 将数据以文本的形式保存到文件中, 可以使用Python的内置函数`open`()来创建、写入和读取文本文件。文本文件是最基本的存储方式, 适合存储结构简单的数据。2. Excel文件是一种常见的存储和处理数据的格式, Python中可以使用`pandas`库来读取、写入和处理Excel文件。`pandas`提供了灵活的数据结构(例如DataFrame)和功能强大的方法来处理和分析数据。3. JSON是一种轻量级的数据交换格式, 常用于存储和传输结构化数据。Python中可以使用`json`模块来处理JSON文件, 可以将JSON数据解析为Python对象, 或将Python对象转换为JSON格式进行存储。

5. 数据预处理

在对网页进行抓拍时, 网页中包含了很多的广告和照片, 并且在抓拍时, 很可能造成数据的扭曲。在Python语言中, 有多种库和工具可供使用来进行数据预处理, 以下是一些常用的数据预处理技术和相应的Python库: 1. 数据清洗: 数据清洗是指处理数据中的错误、缺失值、异常值等问题。Python中的`pandas`库提供了强大的数据清洗功能, 比如删除重复值、填充缺失值、替换异常值等。2. 数据转换: 数据转换是将数据从一种形式转换为另一种形式, 比如数据类型转换、数据编码转换、数据格式转换等。Python中的`pandas`库和`numpy`库提供了丰富的数据转换函数和方法, 可以轻松处理数据的转换需求。利用大数据的分析和处理, 可以达到Python语言下的数据分析和数据处理的程度。

6. 分页检索

首先, 需要连接到数据库, 其次, 编写查询语句, 根据需要的条件和排序规则查询数据, 通过执行查询语句, 可以获取满足条件的全部数据。之后, 可以根据分页的需求, 确定每页显示的数据量和当前页数。当数据量非常大时, 分页检索可能会导致性能问题。为了提高查询效率, 可以考虑使用数据库的索引、优化查询语句等方法。

四、Python语言下大数据分析的设计

根据Python的特点, 可以对其进行数据信息的处理, 使数据阅读的综合程度得到了提升^[2]。在进行数据逻辑分析和信息处理的基础上, 可以利用Python语言来对文件信息和编辑流程等进行优化, 并利用对文件信息的管理和控制, 来提升Python语言在数据分析中的实用

价值。在Python语言建立的文档中，可以使用指令逻辑，对文档中的数据进行处理，在编辑文档中的资料后，可以对文档中的数值运算和登封界面进行优化。首先，需要确定要分析的大数据来源。可以是数据库、日志文件、API接口等。使用Python的第三方库（如Requests、BeautifulSoup等）来获取数据并将其存储在适当的数据结构中，如列表或数据框。其次，对获取的数据进行清洗和预处理，以便后续分析。这包括处理缺失值、处理异常值、数据类型转换等。Python的Pandas库提供了方便的数据清洗和预处理功能。

因此，通过合理运用这些工具和库，可以高效、准确地进行大数据分析，从中发现有价值的洞察和解决方案。并可以对参数和赋值变化等方面展开全面的控制，进而完成对数据的显示和数据的分析，利用cmd指令，对数据进行调用和处理，实现了对数据进行挖掘和信息处理。

五、Python在大数据分析中的基础工具

1. Python程序设计语言介绍

1) Python的特性与优点

1.1 Python简单明了：Python的文法构造清楚明了，容易看懂，减少了程序的复杂度。1.2 海量数据库与模型：Python中包含了海量数据处理、统计分析与机器学习等多个方面的数据库与模型，可为大数据分析与分析提供有力的技术支撑^[3]。1.3 跨平台：Python能够很好地兼容各种应用，便于对各种应用进行数据分析。1.4 有效地开发：Python的开发速度很快，调试速度也很快，这可以帮助数据专家和分析员迅速地建立并对其进行测试。

2. Python类库及工具

1) NumPy及Pandas：DataProcessorMass

1.1 NumPy是Python中的一种非常重要的类库，它为多维阵列的运算和数字的计算提供了有效的手段，同时还可以进行Broadcasting、向定量等运算，加快了对海量数据的处理及计算速度。1.2 Pandas是以NumPy为基础开发的一个数据分析库，它提供了一个高性能的、灵活的、简单的数据结构，以及一些诸如Series、DataFrame等数据的处理方法，为数据的清理、整理和转换提供了便利。

六、结束语

在大数据分析方面，运用Python编程语言，通过本项目的研究，能够有效挖掘大数据，提升大数据分析与信息处理能力。在使用Python的时候，将以xml为格式的输出文件进行阅读，通过网络爬虫的设计和使用，可以提升Python在大数据分析中的实际应用。在信息处理中，包括了文件、执行、脚本和实时信息等内容，利用这些内容，利用这些内容来控制数据输出过程和信息处理等，达到了对输出和相应数据的管理和控制，以适应大数据分析和处理的现实需要。

参考文献：

- [1]李天鸽，徐鹏.大数据专业“Python程序设计”课程建设探索[J].计算机时代，2022（9）：140-142
- [2]陆少兵.基于Python的混合语言程序设计及其实现研究.科技信息，2022（14）：31-33
- [3]张雪莲.Python程序设计语言的特点及应用分析.计算机编程技能与维护，2020（11）：29-30，33