2021年最有用的數據清洗Python 庫

原創 週蘿蔔 蘿蔔大雜燴 2021-12-21 19:29

收錄於話題
#數據清洗 1 #python 28 #數據科學 1 #數據分析師 1

大多數調查表明，數據科學家和數據分析師需要花費**70-80%**的時間來清理和準備數據以進行分析。

對於許多數據工作者來說，數據的清理和準備也往往是他們工作中最不喜歡的部分，因此他們將另外20-30% 的時間花在抱怨上，這雖然是一個玩笑，但是卻很好的反應了數據清洗在數據分析工作當中的特殊地位

在平時的工作生活中，數據總是會出現某些不一致、缺失的輸入、不相關的信息、重複的信息或徹頭徹尾的錯誤等等情況。尤其是當數據來自不同來源時，每個來源都會有自己的一套怪癖、挑戰和不規則之處。凌亂的數據是沒有用處的，有時候甚至會起到反方向作用，這就是數據科學家花費大部分時間來理解所有數據的原因

雖然清理和準備數據既繁瑣又辛苦，但是我們的數據越乾淨、越有條理，後面的一切工作都會變得更快、更輕鬆、更高效。

本文就來分享精選的15 個最有用的Python 數據清理庫，希望在數據分析的路上，大家都能越快輕鬆！

- NumPy

- Pandas
- Matplotlib
- Datacleaner
- Dora
- Seaborn
- Arrow
- Scrubadub
- Tabulate
- Missingno
- Modin
- Ftfy
- SciPy
- Dabl
- Imblearn

NumPy

NumPy 是一個快速且易於使用的開源科學計算Python 庫，它也是數據科學生態系統的基礎庫，因為Pandas 和Matplotlib 等許多最流行的Python 庫都是建立在NumPy 之上的

除了作為其他強大庫的基礎之外，NumPy 還具有許多特性，使其成為Python 數據分析不可或缺的一部分。由於其速度和多功能性，NumPy 的矢量化、索引和廣播概念代表了數組計算的事實標準，NumPy 在處理多維數組時尤為出色。它還提供了一個全面的數值計算工具箱，如線性代數例程、傅立葉變換等

NumPy 可以為很多人做很多事情，其高級語法允許任何背景或經驗水平的程序員使用其強大的數據處理能力。例如，基於NumPy 生成了有史以來第一張黑洞圖像，它還證實了引力波的存在，目前正在各種科學研究中都起著重要的作用

就是這樣一個涵蓋從運動到太空的所有內容的程序也可以幫助我們管理和清理數據，不得不說，Numpy 庫太神奇了

Pandas

Pandas 是由NumPy 提供支持的庫，它是Python 中使用最廣泛的數據分析和操作庫

Pandas 快速且易於使用，其語法非常人性化，再加上其在操作DataFrame 方面擁有令人難以置信的靈活性，使其成為分析、操作和清理數據不可或缺的工具

這個強大的Python 庫不僅可以處理數字數據，還可以處理文本數據和日期數據。它允許我們加入、合併、連接或複制DataFrame，並使用drop() 函數輕鬆添加或刪除列或行

簡而言之，Pandas 結合了速度、易用性和靈活的功能，創建了一個非常強大的工具，使數據操作和分析變得快速而簡單

Matplotlib

了解我們的數據是清理過程的關鍵部分，清理數據的目的就是使其易於理解。但是在我們擁有漂亮乾淨的數據之前，需要先了解凌亂數據中的問題，例如它們的種類和範圍，然後才能對其進行有效的清理，這個操作的很大一部分取決於數據的準確和直觀呈現的程度

Matplotlib 以其令人印象深刻的數據可視化而聞名，這使其成為數據清理工作中的寶貴工具，它是使用Python 生成圖形、圖表和其他2D 數據可視化的首選工具庫

我們可以在數據清理中使用Matplotlib，通過生成分佈圖來幫助我們了解數據的不足之處

Datacleaner

Datacleaner 是一個基於Pandas DataFrame 的第三方庫，雖然Datacleaner 出現的時間比較短並且不如Pandas 流行，但是，Datacleaner 有一種獨特的方法，它結合了一些典型的數據清理功能並使其自動化，這為我們節省了寶貴的時間和精力

使用Datacleaner，我們可以在逐列的基礎上使用眾數或中位數輕鬆替換缺失值，對分類變量進行編碼，並刪除具有缺失值的行

Dora

Dora 庫使用Scikit-learn、Pandas 和Matplotlib 進行探索性分析，或者更具體地說，用於自動化探索性分析中最不受歡迎的方面。除了處理特徵選擇、提取和可視化之外，Dora 還優化和自動化數據清理

Dora 將通過許多數據清理功能為我們節省寶貴的時間和精力，例如輸入缺失值、讀取缺失值和縮放不佳的值的數據以及輸入變量的縮放值等等

此外，Dora 提供了一個簡單的界面，用於在我們轉換數據時保存數據快照，並以其獨特的數據版本控制功能與其他Python 包區別開來

Seaborn

在前面，我們討論了可視化數據以揭示數據缺陷和不一致的重要性。在解決數據中的問題之前，我們需要知道它們是什麼以及它們在哪裡，此時使用數據可視化就是最好的方案。雖然對於許多Python 用戶來說，Matplotlib 是數據可視化的首選庫，然而一些用戶發現Matplotlib 在自定義數據可視化選項方面的局限性也非常大，於是我們有了Seaborn。

Seaborn 是一個數據可視化包，它建立在Matplotlib 之上，可生成有吸引力且信息豐富的統計圖形，同時提供可定制的數據可視化

它也改進了在Pandas 的DataFrames 中的運行效率，可以更加緊密的與Pandas 相結合，使探索性分析和數據清理更加愉快

Arrow

提高數據質量的一個重要方面是在整個DataFrame 中創建統一性和一致性，對於試圖在處理日期和時間時創建統一性的Python 開發人員來說，這個過程可能往往會比較困難。經常在花費了無數個小時和無數行代碼之後，日期和時間格式化的特殊困難仍然存在

Arrow 是一個Python 庫，專門用於處理這些困難並創建數據一致性。它的省時功能包括時區轉換；自動字符串格式化和解析；支持pytz、dateutil 對象、ZoneInfo tzinfo；生成範圍、下限、時間跨度和上限，時間範圍從微秒到數年不等

Arrow 可以識別時區（與標準Python 庫不同），並且默認為UTC。它通過更少的代碼和更少的輸入授予用戶更熟練的日期和時間操作命令。這意味著我們可以為我們的數據帶來更大的一致性，同時減少花在時鐘上的時間

Scrubadub

Scrubadub 是金融和醫療數據科學家的最愛，它是一個Python 庫，專門用於從自由文本中消除個人身份信息(PII)

這個簡單、免費和開源的軟件包可以輕鬆地從我們的數據中刪除敏感的個人信息，從而保護當事人的隱私和安全

Scrubadub 目前允許用戶清除以下信息的數據：

- 電子郵件地址
- 網址

- 姓名
- Skype 用戶名
- 電話號碼
- 密碼/用戶名組合
- 社會安全號碼

Tabulate

只需調用一個函數，Tabulate 就可以使用我們的數據創建小型且有吸引力的表格，由於具有數字格式、標題和小數列對齊等許多功能，這些表格具有很高的可讀性

這個開源庫還允許用戶使用其他工具和語言處理表格數據，讓用戶能夠以其他擅長的格式（如HTML、PHP 或Markdown Extra）輸出數據

Missingno

處理缺失值是數據清理的主要方面之一，Missingno 庫應運而生。它逐列識別和可視化DataFrame 中的缺失值，以便用戶可以看到他們數據所處的狀態

將問題可視化是解決問題的第一步，而Missingno 是一個簡單易用的庫，可以很好的完成這項工作

Modin

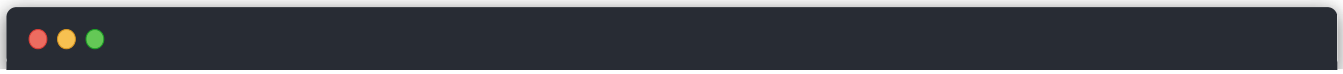
正如我們上面提到的，Pandas 已經是一個快速的庫了，但Modin 將Pandas 帶到一個全新的水平。Modin 通過分發數據和計算速度來提高Pandas 的性能

Modin 用戶將受益於與Pandas 語法的完美契合和不顯眼的集成，可以將Pandas 的速度提高多達400%！

Ftfy

Ftfy 的誕生是為了一個簡單的任務：將糟糕的Unicode 和無用的字符轉換為相關且可讀的文本數據

比如：



```
â€œquoteâ€\x9d = "quote"
uÌˆ = ü
lt;3 = <3
```

無需花費大量時間處理文本數據，使用Ftfy 就可以快速理解無意義的內容

SciPy

SciPy 不僅僅是一個庫，它還是一個完整的數據科學生態系統

此外，SciPy 還提供了許多專用工具，其中之一是Scikit-learn，完美可以利用其“Preprocessing”包進行數據清理和數據集標準化

Dabl

scikit-learn 項目的一名核心工程師開發了Dabl 作為數據分析庫，以簡化數據探索和預處理的過程

Dabl 有一個完整的流程來檢測數據集中的某些數據類型和質量問題，並自動應用適當的預處理程序

它可以處理缺失值，將分類變量轉換為數值，它甚至具有內置的可視化選項以促進快速數據探索

Imblearn

我們要介紹的最後一個庫是Imbalanced-learn（縮寫為Imblearn），它依賴於Scikit-learn 並為面臨分類和不平衡類的Python 用戶提供工具支持

使用稱為“undersampling”的預處理技術，Imblearn 將梳理完美的數據並刪除數據集中的缺失、不一致或其他不規則數據

總結

我們的數據分析模型取決於我們輸入的數據，並且我們的數據越乾淨，處理、分析和可視化就越簡單，善於利用工具，會使我們的工作更加輕鬆愉快

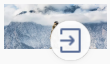
雖然上面總結的工具不可能包含所有的數據清洗工具，但是我們只要選擇適合我們的就可以了，希望今天的分享能夠幫助到你~

好了，今天分享就到這裡，如果大家覺得滿意請務必點個**贊**+**在看** 支持下

往期推薦

太好玩了，爬蟲、部署API、加小程序，一條龍玩轉知乎熱榜！

2021-12-16



使用Python自動製作《歷史上的今天》宣傳圖片

2021-12-15

Python 教你自動發微博，每日一句英語

2021-12-14