Installationen für Data Science mit Python

NumPy Übersicht

- NumPy (oder Numpy) ist eine Library für lineare Algebra für Python
- Sie ist deshalb für Data Science mit Python so wichtig, da fast alle anderen Libraries im Python Ökosystem auf NumPy aufbauen
- NumPy ist außerdem extrem schnell, da es Verknüpfungen zu C hat

Installation

- Die Installation von NumPy (und anderen Libraries) funktioniert am einfachsten und zuverlässigsten, wenn wir die im Kurs empfohlene Anaconda Distribution von Python nutzen
- Bei der Nutzung von Anaconda, installiere Numpy im Terminal bzw. der Kommandozeile durch folgenden Befehl:

```
conda install numpy
```

```
pip install numpy
```

NumPy im unserem Kurs

DATAMICS

- NumPy Arrays sind der primäre Weg, wie wir NumPy im Verlauf dieses Kurses nutzen werden
- NumPy Arrays gibt es hauptsächlich in zwei Varianten: Vektoren und Matrizen
- Vektoren sind streng eindimensionale Arrays (1-d), wohingegen Matrizen zweidimensionale Arrays (2-d) sind. Dabei gilt zu beachten, dass auch Matrizen nur eine Spalte bzw. nur eine Zeile haben können.

Pandas Übersicht

- Pandas ist eine Open Source Library, die auf NumPy aufgebaut ist
- Es erlaubt uns schnelle Analysen und Datenbereinigung sowie -vorbereitung
- Es zeichnet sich besonders durch Performance und Produktivität aus
- Zusätzlich beinhaltet es vorinstallierte Visualisierungsfunktionalität
- Pandas kann mit einer Vielzahl an Datenquellen arbeiten

Installation

- Wir müssen Pandas (wie die meisten anderen Libraries im Verlauf des Kurses) installieren.
- Dazu gehen wir ins Terminal bzw. die Kommandozeile und führen folgendes aus:

```
conda install pandas
```

```
pip install pandas
```

Datenquellen für Input

- CSV
- Excel
- HTML
- SQL

Module

- Dazu benötigen wir die folgenden Module die wir mit pip oder conda wie folgt installieren können:
  - *conda install sqlalchemy*
  - *conda install lxml*
  - *conda install html5lib*
  - *conda install BeauytifulSoup4*

MatplotLib

- Wir haben jetzt einen guten Überblick über die Datenanalyse mit NumPy, Pandas und SciPy
- Das bildet die Grundlage die wir brauchen um fortzuschreiten
- Als nächstes widmen wir uns der Visualisierung von Daten
  - MatplotLib
  - Seaborn
  - Pandas vorinstalliere Visualisierung
  - Plotly für interaktive Plots

MatplotLib Übersicht

DATAMICS

- MatplotLib ist die bekannteste und am meisten verbreitete Library zum Darstellen von Diagrammen und ähnlichen Visualisierungen
- Es erlaubt uns jeden einzelnen Aspekt einer Darstellung einzustellen
- Die Idee bei der Erstellung war es eine ähnliche Nutzererfahrung zu bieten, wie es MatLab‘s Darstellung tut

Installation

- Zur Installation von MatplotLib gehen wir ins Terminal bzw. die Kommandozeile und geben folgenden Befehl ein:

```
conda install matplotlib
```

```
pip install matplotlib
```

Seaborn

- Seaborn ist eine statistische Plotting Bibliothek um Diagramme darzustellen
- Standardmäßig beinhaltet es schon sehr schöne Styles und Diagramme
- Es harmoniert sehr gut mit den Pandas DataFrame Objekten

Installation

- Zur Installation von Seaborn gehen wir ins Terminal bzw. die Kommandozeile und geben folgenden Befehl ein:

```
conda install seaborn
```

oder

```
pip install seaborn
```

Plotly und Cufflinks

- Plotly ist eine interaktive Visualisierungs-Library
- Cufflinks verbindet diese Library mit Pandas
- Bevor wir diese nutzen können müssen wir beide installieren

Installation

- Zur Installation von Plotly und Cufflinks können wir Anaconda leider nicht direkt verwenden.
- Wir gehen wir ins Terminal bzw. die Kommandozeile und geben folgenden Befehl ein:

```
pip install plotly
```

und

```
pip install cufflinks
```

Exklusive Gutscheine

DATAMICS

Verwende den Gutschein „**SLIDESHARE2018**“ auf *Udemy* oder die Shortlinks und erhalte unsere Kurse für nur 10,99€ (95% Rabatt).

Python für Data Science und Machine Learning: https://goo.gl/cE7TQ3

Original Python Bootcamp - Von 0 auf 100: https://goo.gl/gjn7pX

R für Data Science und Machine Learning: https://goo.gl/8h5tH7