Hochschule Bochum
Bochum University
of Applied Sciences

Fachbereich Wirtschaft
Prof. Dr. Christian Bockermann

Data Science

Tutorial Session 1

Diese Aufgabenblätter dienen als zusätzliche Übung. Die Aufgaben können jederzeit bearbeitet werden.

Aufgabe 1 (Strings und Dokumente)

In dieser Aufgabe betrachten wir Text-Dokumente. Ein Dokument ist dabei erstmal in einer einfachen Form als `str` Objekt gegeben, z.B.:

```
dokument1 = "Python macht Spass!"
```

Der Datentyp `str` bietet eine Reihe hilfreicher Funktionen an, z.B:

```
# replace:
dokument2 = dokument1.replace("Python", "Programmieren")

# split
woerter = dokument2.split()
```

1. Erzeugen Sie die Variable `dokument3`, in dem Sie `dokument1` und die `replace()` Funktion benutzen und `Python` durch `DataScience` ersetzen!

2. Welchen Typ liefert die Funktion `split()` (siehe obiges Beispiel) zurück?

3. Definieren Sie eine Funktion `anzahl(b, s)`, die für einen Buchstaben `b` und einen String `s` zurückgibt, wie häufig der Buchstabe `b` in `s` enthalten ist.

4. Definieren Sie eine Funktion `buchstaben(s)`, die für einen String die *Menge* der Buchstaben des Strings zurückgibt.

   Hinweis: Mit *Menge* ist hier natürlich gemeint, dass jeder Buchstabe nur *einmal* enthalten ist. Mengen werden in Python durch den Typ `set` dargestellt, der sich aus einer Liste erzeugen lässt:

   ```
   xs = [1, 1, 4, 2, 2, 5]
   ys = set(xs)   # ys ist dann:  {1, 4, 2, 5}
   ```

5. Schreiben Sie eine Funktion `bsAnzahl(s)`, die für einen String eine Liste von Paaren (2er-Tuple) zurückgibt, die jeweils den Buchstaben und die Anzahl enthält, also:

   ```
   s = "AaabbbBB"
   xs = bsAnzahl(s)
   # xs ist dann  [('A',1), ('a',2), ('b', 3), ('B', 2)]
   ```

6. Schreiben Sie eine Funktion `uniCount(xs)`, die für eine Liste von Tupel aus Buchstaben und Anzahl, die Summe der Anzahlen berechnet. Im Falle der vorherigen Aufgabe ist das Ergebnis also 1 + 2 + 3 + 2 = 8.

Aufgabe 2 (Pandas)

Unter der URL

`https://data.hsbo.de/wetter_basel.csv`

finden Sie einen Datensatz mit Wetterdaten aus Basel. Die Daten sind ein Export seit dem 1.1.2001.

1. Laden Sie den Datensatz in einen DataFrame und inspizieren Sie die Daten. (Denken Sie daran, dass Sie das Pandas Modul importieren müssen!)

   Wieviele Datensätze enthält der Datensatz? Wieviele Attribute?
   Lassen Sie sich den DataFrame im Notebook anzeigen!

2. Rufen Sie im Notebook den Befehl `help(pd.DataFrame.dropna)` auf und finden Sie heraus, was die Funktion tut.

   Wenden Sie `dropna()` auf den DataFrame an und schauen Sie sich das Ergebnis an!

3. Dummerweise ist die Temperatur in Fahrenheit angegeben. Googlen Sie z.B. nach "fahrenheit to celsius" um eine Umrechnungsformel zu finden.

   Schreiben Sie eine Funktion `f_to_c(x)`, die einen Wert `x` von Grad Fahrenheit in Grad Celsius umrechnet!

4. Mit `s.apply(f)`, lässt sich eine Funktion `f` auf jedes Element einer Series `s` anwenden. Das Resultat ist eine neue Series mit den elementweisen Ergebnissen.

   Nutzen Sie `s.apply(..)` um eine Spalte mit Celsius-Temperaturen zu ihrem DataFrame hinzuzufügen.

5. Mit der Funktion `df.plot(y="Spalte")` lässt Sich eine Spalte des DataFrames plotten. Erzeugen Sie einen Plot für die Temperatur-Spalte!

   Der Parameter `y` in der Plot-Funktion erlaubt auch eine Liste von Spaltennamen. Erzeugen Sie einen Plot für die Temperatur-Spalten mit Celsius und Fahrenheit!

6. Die Spalte `timestamp` ist vom Typ `str`. Um daraus ein richtiges Datum zu erzeugen, bietet Pandas die Funktion `pd.to_datetime(..)` an. Wenn die Funktion `pd.to_datetime(s)` eine Series `s` als Parameter bekommt, ist das Ergebnis eine Series mit Datums-Objekten.

   Berechnen Sie eine Spalte mit Datums Objekten!

   Ersetzen Sie den **index** ihres DataFrames durch die neue Datumsspalte und schauen Sie sich den DataFrame an. Erzeugen Sie die Plots für die Temperaturen nochmal (am besten in einer neuen Zelle) und vergleichen Sie die Plots mit den vorherigen Plots!