Merkblatt CSV

1 Comma Separated Values (CSV)

CSV steht für Comma Separated Values, also Komma-separierte Werte. Das bezeichnet ein gängiges Format zum Austausch von Werten in Listen oder sogar Datenbanken. Es gibt zwar einen Standard für CSV (RFC 4180), aber dieser ist nicht verbindlich und das CSV-Format in Variationen ist auch schon länger verbreitet. Daher gibt es verschiedene Varianten.

Die wesentliche Unterschiede in der Beschreibung sind zum einen das Trennzeichen. Zuweilen gibt es Formate, die anstatt des namensgebenden Kommas andere Zeichen verwenden: Semikolon, Tabulatoren oder sogar nur Leerzeichen. Da diese Werte selber in Zeichenketten vorkommen können, werden zum anderen alle Daten, Zeichenkette oder nur Zeichenketten mit eben diesen Zeichen in Anführungszeichen gespeichert. Hinzu kann noch ein Escape-Zeichen kommen - meist wie in Python ein.

Kombiniert mit dem Herausfinden der richtigen Zeichenkodierung einer Datei unbekannten Exports kann dies einige Zeit an Probierne bedeuten. Ähnlich dem Importiermodul für CSV von MS Excel und LibreOffice kann auch in Python die Import-Fuktion der Panda-Bibliothek mit vielen Parameter eingestellt wwerden.

1.1 CSV importieren - pd.read_csv()

Um CSV-Dateien zu importieren, muss zunächst das Pandas Modul mit folgender Codezeile importiert werden. Es ist Konvention, das Pandas-Modul dabei nachfolgend als `pd` anzusprechen.

```
import pandas as pd
```

Dann wird die die Datei (hier `CSV-Datei.csv`) importiert und als Dataframe (hier `df`, ebenfalls eine als Konvention verwendete Bezeichnung) abgespeichert.

```
df = pd.read_csv("CSV-Datei.csv")
```

1.1.1 Einstellungen zum CSV-Import

Um eine kompriemierte CSV-Datei zu verwenden, muss sie nach dem Einlesen noch entpackt werden. Python erledigt dies automatisch, wenn die Dateiendung bereits auf den Kompressionsalgorithmus hinweist (hier `.bz2`)

```
df = pd.read_csv("CSV-Datei.csv.bz2")
```

Unter anderem folgende Formate bzw. Dateiendungen sind gängig: .bz2, .gz, .lzma, .tar, .zip, .xz Für einige Kompressionsmethoden muss zunächst eine Bibliothek importiert werden. Zu weitern Informationen hilft die Dokumentation und Google.

1.2 Einige Parameter

Mit einer Vielzahl an Parametern kann der CSV-Import gesteuert werden, falls es mit der selbständigen Erkennung in Python oder Standardeinstellungen nicht funktioniert. Hierzu sei auch auf die Dokumentation verwiesen.

Die wichtigsten Parameter folgen:

1.2.1 sep or delimiter

Der wichtigste Parameter ist die Einstellung des Trennzeichens zwischen zwei Zellen. Dieses Zeichen wird auch als Separator bezeichnet. In der Regel sollte es sich um ein Komma (,), ein Semikolon (;) oder einen Tabulator (⁀) handeln. Mit `sep="\t"` wird zum Beispiel der Tabulator als Trennzeichen eingestellt. Andere Zeichen sind auch möglich.

Der Parameter `delimiter` ist eine andere Bezeichnung für den gleichen Parameter.

1.2.2 compression

Der Kompressionsalgorithmus kann explizit mit dem Parameter `compression` angegeben werden. Dadurch wird auch ggf. die Kompressionsmethode entsprechend der Dateiendung übergangen. Die bekannte Datei `CSV-Datei.csv.bz2` könnte auch so eingelesen werden

```
df = pd.read_csv("CSV-Datei.csv.bz2", compression="bz2")
```

1.2.3 decimal

Der Dezimalpunkt ist standardmäßig ein Punkt. Das entspricht dem US-amerikanischen Standard. mit `decimal=","` wird dies in das "deutsche" Komma geändert.

1.2.4 thousands

Bei der US-amerikanischen Notation sind Punkt und Komma auch hinsichtlich der Schreibweise von Tausendern getauscht. "1,000" ist dann Eintausend. Mit dem Parameter `thousands` kann dies ebenso verändert werden.

1.2.5 quotechar

Zeichenketten sind in einer CSV-Datei ggf. in Anführungszeichen gesetzt. Wird das in Python nicht automatisch richtig erkannt, so kann die Art dieses Anführungszeichen mit `quotechar="'"` angegeben werden. Hier würde explizit das einfache Hochkomma verwendet.

1.2.6 lineterminator

Normalerweise endet eine Zeile mit einem Zeilenumbruch `\n`. Das kann aber auch anders sein. Es können zwei Zeilenumbrüche sein `\n\n`. Auch jedes andere Zeichen kommmt infrage.

[ ]: