

Informatik für Data Science 2

Tutorium - Termin 3 (12.04.2021)

Die Lösung (und weiteres Material) sind in GitHub (Repository: <https://github.com/hka-mmvmv/dscb230-tutorial.git>)

In dem Berufsfeld des Data Scientist solltet ihr in der Lage sein anhand von Daten wissen zu generieren. Dies *kann* zwei Schritte benötigen. Zunächst müssen die Daten in eine angebrachte Form gebracht werden und gegebenenfalls etwas bearbeitet. Im zweiten Schritt werden diese dann analysiert und wissen generiert. Im heutigen Termin wird der erste Schritt geübt. In Github befinden sich neben dem Template auch noch zwei weitere Json Files. Diese Datensätze werden anhand den folgenden Teilaufgaben verarbeitet.

*Im folgenden wird der Datensatz T_20210412_data_1.json als Datensatz 1 betitelt.
Im folgenden wird der Datensatz T_20210412_data_2.json als Datensatz 2 betitelt.*

1. Die Datensätze wurde für euch bereits eingelesen und in Klassen Variablen gespeichert. Findet nun ein gemeinsames Attribut der beiden Datensätze. Damit solltet ihr nun in der Lage sein die beiden Datensätze zusammenzuführen. Bitte erstellt hierfür ein neues Attribut `more` welches ein Dictionary mit den entsprechenden Daten aus Datensatz 2 enthält. Schreibt hierfür eine Methode `expand_data()` welches genau diese Aufgabe übernimmt. Das Ergebnis sollte etwa so aussehen:

```
[
  {
    "id": 1,
    "street_name": "692-7375 Vestibulum Rd.",
    "postal_code": "233730",
    "city": "Brugge Bruges",
    "country": "Cape Verde",
    "location": "-85.23667, -2.31575",
    "more": {
      "id": 1,
      "name": "Thompson, Teagan Q.",
      "email": "risus.Donec.nibh@In.net",
      "personalnumber": "16900108 0200"
    }
  },
  ...
]
```

2. Bitte findet nun logisch ähnliche Attribute aus Datensatz 1 und fügt diese zusammen. Beispielsweise können `city` und `country` zusammen geführt werden, also ein Tupel aus Straße und Hausnummer. Schreibt hierfür eine Methode `fuse_related_values()` welches genau diese Aufgabe übernimmt. Das Ergebnis sollte etwa so aussehen:

```
[
  {
    "id": 1,
    "street_name": "692-7375 Vestibulum Rd.",
    "city": ["233730", "Brugge Bruges"],
    "country": ["Cape Verde", "-85.23667", "-2.31575"],
    "more": {
      "id": 1,
      "name": "Thompson, Teagan Q.",
      "email": "risus.Donec.nibh@In.net",
      "personalnumber": "16900108 0200"
    }
  },
  ...
]
```

3. Zuletzt soll jedes einzelne Dictionary mit einem Hash Key versehen werden. Hierfür könnt ihr ein beliebiges aber eindeutiges Element aus dem Datensatz wählen. Schreibt hierfür eine Methode `hash_keys()` welches genau diese Aufgabe übernimmt. Das Ergebnis sollte etwa so aussehen:

```
{
  "17aaa25daf5289b4a07d3b26f73ec1bcd1fc530fb09c63b76cb3dba3d974f47b": {
    "id": 1,
    "street_name": "692-7375 Vestibulum Rd.",
    "city": ["233730", "Brugge Bruges"],
    "country": ["Cape Verde", "-85.23667", "-2.31575"],
    "more": {
      "id": 1,
      "name": "Thompson, Teagan Q.",
      "email": "risus.Donec.nibh@In.net",
      "personalnumber": "16900108 0200"
    }
  },
  ...
}
```

Nun habt ihr eine beispielhafte Datenverarbeitung erfolgreich erarbeitet. Bei unserem nächsten Termin werden wir die Daten Graphisch darstellen und analysieren.