Hochschule Bochum
Bochum University
of Applied Sciences

Fachbereich Wirtschaft
Prof. Dr. Christian Bockermann

Data Science 2

Wintersemester 2021/2022

Übungsblatt 5

Aufgabe 1 (Häufige Mengen)

Bei der Warenkorbanalyse geht es ja darum, die häufigen Mengen auf einer Transaktionsdatenbank zu berechnen und daraus Assoziationsregeln abzuleiten. In der Datei

`Kurse/DataScience2/data/transactions.csv`

finden Sie bereits einen Datensatz, der die Daten als Transaktionsdatenbank enthält. In dieser Aufgaben sollen Sie zunächst die häufigen Mengen berechnen und dann die Assoziationsregeln daraus erzeugen.

1. Laden Sie die Daten in einen `DataFrame`.

   Wie viele Transaktionen enthält der Datensatz?
   Wie groß ist die Anzahl der Produkte?

2. Wie groß (Anzahl der Produkte) sind die Transaktionen im Durchschnitt?
   Wieviele Produkte enthalten die Transaktionen maximal?

   Hinweis: Schauen Sie sich die Hilfe zur Funktion `sum(..)` von DataFrame an – insbesondere den Parameter `axis`.

3. Berechnen Sie die häufigen Mengen mit der Funktion `apriori` aus dem Modul `mlxtend.frequent_patterns`. Probieren Sie dafür einen minimalen Support von 0.1 und später 0.05 aus.

   Wieviele Mengen werden für den jeweiligen minimalen Support Wert gefunden?

   Hinweis: Der Parameter `min_support` legt den minimalen Support fest. Für eine leserliche Darstellung empfiehlt es sich auch den Parameter `use_colnames=True` zu nutzen.
   Ein Beispiel dazu finden Sie auf Folie 18.

4. Die häufigen Mengen werden von `apriori` wieder als DataFrame Objekt ausgegeben. Die Spalte `itemsets` enthält dabei die *Liste* der Symbole/Produkte.

   Filtern Sie alle Mengen heraus, die mindestens 2 Symbole/Produkte enthalten.

5. Benutzen Sie die Funktion `association_rules` aus dem gleichen Modul um die Assoziationsregeln zu berechnen.

Aufgabe 2 * (Transaktionsdatenbank)

In der Regel liegen die Daten nicht vorab als Transaktionsdatenbank vor. Die Rohdaten aus Aufgabe 1 haben z.B. das Format

```
avocado, almonds, shrimp
almonds, flour, olive, shrimp
tea, olive
```

wobei jede Zeile eine Transaktion mit den enthaltenen Produkten darstellt. Die Rohdaten finden Sie in der Datei

`Kurse/DataScience2/data/baskets.csv`

In dieser Aufgabe geht es darum, die Daten so aufzubereiten, dass daraus eine Transaktionstabelle wird, die wir für z.B. den Apriori Algorithmus verwenden können (vgl. Aufgabe 1). Das heisst, wir benötigen einen DataFrame der folgenden Art:

ID	avocado	almonds	tea	flour	olive	shrimp
1	1	1	0	0	0	1
2	0	1	0	1	1	1
3	0	0	1	0	1	0
4	0	1	0	0	1	1
5	1	1	0	0	1	0

Es gibt verschiedene Wege, zu solch einer Darstellung zu kommen. Ein Ansätze könnte z.B. sein

- Daten in DataFrame lesen, dabei nur 1 Spalte mit den Artikeln
- Aus der Text-Spalte mit `split` Wort-Listen erzeugen
- Aus Wort-Listen dann mit `CountVectorizer` Word-Vector Darstellung erzeugen (vgl. DataScience 2, Vorlesung 4, Folie 26)

Berechnen Sie aus den Rohdaten eine Transaktionstabelle als DataFrame!