Data Science 2

Übungsblatt 4

Aufgabe 1 (Clustering)

In Data Science 1 hatten wir Daten zu einem Versandunternehmen betrachtet, bei dem Sendungen/Pakete nach Größe, Gewicht und Volumen von unterschiedlichen Logistik-Dienstleistern transportiert wurden. Die Daten finden Sich in der Datei

`https://data.hsbo.de/versand-data.csv`

In dieser Aufgabe sollen die Sendungen gruppiert werden.

1. Laden Sie die Daten in einen DataFrame und verschaffen Sie sich einen Überblick.

   Was für Attribute gibt es? Welchen Typ haben diese?

2. Berechnen Sie mit dem k-Means Algorithmus ein Clustering auf den Daten. Starten Sie dazu mit $k = 5$ Clustern. Jedes Cluster ist also eine Teilmenge der Sendungen.

3. Schreiben Sie eine Funktion `gewichte(..)`, die für einen Cluster die Häufigkeiten der Logistik-Dienstleister berechnet, die in dem Cluster vertreten sind, also z.B. ein Ergebnis der Art:

   ```
   gewichte(cluster0) = { 'UPS': 43, 'DHL': 17, 'WPS': 6 }
   ```

   Hinweis: Bedenken Sie, dass k-Means nur mit numerischen Spalten funktioniert. D.h. Sie müssen die `Dienstleister`-Spalte für das Clustering ausblenden und später wieder hinzufügen.

4. Benutzen Sie das Modul `wordcloud` um eine WordCloud mit den berechneten Gewichten eines Cluster zu berechnen und anzuzeigen.

   Das `wordcloud` Modul ermöglicht die Benutzung von eigenen Gewichten auf die folgende Weise:

   ```
   from wordcloud import WordCloud
   import matplotlib.pyplot as plt

   wc0 = WordCloud()
   wc0.generate_from_frequencies( gewichte(cluster0) )

   plt.imshow(wc0)
   ```

5. Schreiben Sie ein Funktion `wolke(..)`, die für einen Cluster die WordCloud berechnet. Erzeugen Sie für jeden ihrer Cluster eine WordCloud.

6. Probieren Sie ihr Clustering für $k = 2$, $k = 3$ und $k = 4$ aus. Wie verändern sich die WordClouds?