Aufgabenblatt 3

1. Verwenden Sie DBSCAN, um Clusters von nah aneinanderliegenden großen Städten der Erde (Ballungsräume) zu finden. Eine Stadt mit mindestens 50.000 Einwohnern gilt als groß. Die ϵ-Nachbarschaft einer Stadt enthält alle anliegenden Städte mit einem Euklidischen Abstand von höchstens 0,15 in Bezug auf Breiten- und Längengrad. Eine Stadt gilt als Kernobjekt eines Ballungsraums, wenn mindestens 8 Städte in ihrer ϵ-Nachbarschaft liegen. Verwenden Sie zum Clustering den Datensatz `maps::world.cities`. Beantworten Sie die folgenden Fragen:

   a) Wie viele Clusters, Kernobjekte, Randobjekte und Noise-Objekte werden von DBSCAN gefunden?
   b) Wie viele Städte beinhaltet das größte Cluster und in welchem Land liegen die Städte des größten Clusters?
   c) Welche drei Ländern verfügen über die meisten Städte in Clusters?
   d) Sind die indischen Städte `Rajendranagar` und `Rajpur` (direkt) dichte-erreichbar oder dichte-verbunden?
   e) Sind `Essen` und `Castrop-Rauxel` (direkt) dichte-erreichbar oder dichte-verbunden?
   f) Welche Städte sind von `Bochum` aus dichte-erreichbar, aber nicht *direkt* dichte-erreichbar?

2. Gegeben Sei erneut der Datensatz aus Aufgabe 2 des Aufgabenblatts 2. Verwenden Sie dieses Mal zum Clustering DBSCAN mit $minPts = 6$. Bestimmen Sie zunächst einen *geeigneten* Wert für ϵ. Stellen Sie das Clustering in einem Scatter Plot dar. Heben Sie Clusterzuordnungen und Ausreißer (Noise Points) farblich hervor. Vergleichen und diskutieren Sie das Clustering von DBSCAN mit dem Clustering von k-Means.

3. Gegeben Sei erneut der Datensatz aus Aufgabe 2 des Aufgabenblatts 2. Verwenden Sie OPTICS, um ein Dichte-Erreichbarkeitsdiagramm für $minPts = 6$ zu erstellen. Extrahieren Sie jeweils ein Clustering für *reachability-dist* $= \{1, 1.5, \ldots, 5\}$ und stellen Sie das Ergebnis jeweils in einem Scatter Plot dar. Heben Sie Clusterzuordnungen und Ausreißer (Noise Points) farblich hervor. Bewerten Sie die Veränderung des Clustering-Ergebnisses mit zunehmenden Schwellwert für *reachability-dist* bezüglich der Anzahl von Clusters sowie der Anzahl von Core Border, und Noise Points.

4. Diskutieren Sie am Beispiel des Silhouettenkoeffizienten die Stärken und Schwächen von internen Qualitätsmaßen? Warum sind sie für den Vergleich zwischen Clusterings verschiedener Algorithmen (z.B. K-Means und DBSCAN) nur bedingt geeignet? In welchen Fällen sollte man sie dennoch einsetzen?

Datensatz für Aufgaben 2 und 3:
http://isgwww.cs.uni-magdeburg.de/cv/lehre/VisAnalytics/material/exercise/datasets/clustering-student-mat.csv