Exercise 5

Exercise 5 is due to 20.6.2017 (uploaded on CEWebS) and will be discussed on 21.6.2017.

1. Temporal Analysis

The data on timeseries.xlsx show the monthly access of to a server from users in two countries and different locations. Additionally, the total accesses from the two countries are shown.

The following questions are of interest:

a) Are there differences in the patterns in the locations over the period of three years?
b) In how far show the patterns on the different locations similarities with the overall pattern of the countries?
c) Are there substantial differences in the access patterns between the two countries?
d) Find simple summary characteristics for the access patterns.

Use visualization techniques for the series and use time warping for defining distances between the access behaviors. In how far differ the results using summary characteristics from the results using the summary measures.

2. Text mining

Data preprocessing, quality assessment, and data warehousing are important activities in all BI projects. Both topics have not only the computer science aspect but also more statistical oriented aspects. For example, it is important to know the provenance of the data, the reference universe of the data and information about the different methods applied for improving data quality.

In the document archives `DataProcessing.7z` and `DWH.7z` are a number of documents dealing with these aspects for a statistical data warehouse.

The documents were downloaded from the projects *Memobust* and *Data Warehouse* from the CROS portal of Eurostat (http://www.cros-portal.eu/content/finished-projects) which is Collaboration platform for Research in Official statistics. (The documents are public available without registration.)

The main task is to apply text mining to identify interesting topics and keywords, assignment of the different documents to the keywords and topics.

Using this approach one can think about a more user friendly organization of the material and compare this approach with the more computer science oriented approaches towards data warehousing and data quality.