Project 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 processing 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 `processing.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.