Developoing textmining system for the Hungarian Electronic Library

The essay describes three aspects of the textmining system developed for the Hungarian Electronic Library (HEL) supported by Infopark Alapítvány. First, the system architecture is provided in detailes which is a scalable and distributed client-server solution based on simple object access protocol communication. The applied compact document reprezentation method is described that is able to store HTML and XML documents losslessly where the textmining functions can be performed in this compact reprezentation with high performance. This reprezentation is based on a special database structure indexed by a contributing finite state automaton. Finally, the main textmining functions are described from the viewpoint of how they can meet the needs of HEL and the commercial entities.