Knowledge mining from PDF's

Siemens is a global powerhouse focusing on the areas of electrification, automation and digitalization. One of the world's largest producers of energy-efficient, resource-saving technologies, Siemens is a leading supplier of systems for power generation and transmission as well as medical diagnosis. In infrastructure and industry solutions the company plays a pioneering role. As of September 30, 2014, we had around 343,000 employees in more than 200 countries.

Siemens Industrial Turbomachinery AB (SIT AB) in Sweden is part of the Siemens Energy Sector. The Energy Sector is the world's leading supplier of products, services and solutions for the generation, transmission and distribution of power and for the extraction, conversion and transport of oil and gas. SIT AB delivers gas turbines, steam turbines, turn-key power plants, service and components for heat and power production. All under one roof – from research and development, manufacturing, marketing, sales and installation of turbines and complete power plants to service and refurbishing. There are today about 2 700 employees in Finspång.

Project Field Experience In SIT AB, a large amount of field experience data is continuously generated in form of various reports from maintenance events, component repair and operation history. These reports include detailed information about the turbine operation history as well as its condition and reported damages on individual components. This field experience data, although noisy, invariably portray environmental factors, measurement errors, and loading conditions, or in short, reality. By establishment of a process to collect and maintain this information in a database format, exploration and knowledge discovery using this data became a subject of high interest. This Master thesis is a part of efforts done to improve the data feeding part of the project by automating information capturing.

Project description

Manually digitalizing data from .pdf documents is not innovative and requires time and effort. If the whole or some part of this task could be digitalized, great gains in time and effort would be made.

This Master Thesis project consists of three parts:

I. Literature survey of relevant techniques for capturing data from .pdfs
II. Implementing a text-mining tool based on technique chosen in I. above on e.g. 5000 .pdf files along with a set of relevant information to be retrieved.

The project is suitable for 1-2 students with good background in text mining. Student will work closely with domain experts.

For further information, please contact Daniel Dagnelund, daniel.dagnelund<at>siemens.com, or Davood Naderi davood.naderi<at>siemens.com.