PDF/graphic file analysis tool sought for conversion to machine-readable format

-
- Scheda
- Approfondimenti

PDF/graphic file analysis tool sought for conversion to machine-readable format

Identificativo proposta:TRAT20190926002 **RICHIEDI MAGGIORI INFORMAZIONI**

A small Austrian engineering company seeks a solution, which can extract the elements of a given complexly structured document (primarily PDF but also other formats) and convert them into a machine-readable form. Elements include text, graphics, drawings, tables, headlines and info-boxes. Very high accuracy is needed (90-99% correct readings). The company seeks partners for licence agreement, commercial agreement with technical assistance or research cooperation agreement.

An Austrian engineering firm is looking for a solution to facilitate information processing and the analysis of data particularly in PDF and graphic format. When for example researching a topic online, most information is contained in free formatted documents, which contain a number of different elements. In order to analyse the documents, they first need to be broken down into formal elements such as text, images, headings, tables etc. and the logical links need to be retained. Only then can the contents be interpreted by specialised, external services, e.g. NLU (natural language understanding) or image recognition. While there are a number of solutions currently available, the one major shortcoming is their insufficient extraction hit rate. The offered solution must deliver accuracy of 90-99% depending on the text complexity. Current solutions are also very focused on plain text. Context information (headings, text placement, to which graphic does it refer), tables, multiple columns of the text, graphics, images, captions or comments cannot be reliably extracted in context. The Austrian SME is looking for a solution to handle input primarily in the form of good quality PDF scans. However, a tool that can also deal with Word, Excel, PowerPoint and .jpg formats would be ideal. It is important that the solution can handle complex content with headlines, sublines, different text bodies, columns of text, drawings, graphics, tables, nesting tables, info boxes (text) and sometimes background pictures. The tool should cover the following processing steps: • breaking down a given document into the elements mentioned above. Context-information should be retained (headline of text body, page number, elements close to each other etc.); • converting text, numbers and tables into machine-readable formats; • displaying confidentiality-index of each separate element described in the section "technical specification or expertise sought" • outputting the results in file systems or databases. The technical specifications for the initial solution are detailed in the section "technical specification or expertise sought". Future versions of the solution, which can handle tables of contents, non-Latin characters (Chinese, Japanese, Korean) and more complex background images are also of interest. The ideal cooperation types include: - a licence agreement where a vendor is willing to sell a licence for a suitable solution; - a commercial agreement with technical assitance with a developer who can deliver the required framework with complete knowledge transfer. In this case, initial training should be provided by the developer to an extent that (almost) no further training is required; - a research cooperation agreement, where the Austrian company would work together with another party to develop the technology, should a suitable company or research team be found.

Riferimento Esterno: TRAT20190926002

Tipo: Technology Request

Paese: Austria

Presentazione: 09/10/2019

Ultimo aggiornamento: 15/10/2019

Scadenza: 15/10/2020