Karlsruher Institut für Technologie

Studienzentrum für Sehgeschädigte (SZS)

Masterarbeit: Erkennung mathematischer Formeln in PDF-Dokumenten

Das Studienzentrum für Sehgeschädigte ist immer an neuen assistiven Zugängen interessiert, um es blinden Menschen im Alltag bzw. betroffenen Studierenden im Studienalltag einfacher zu machen.

In elektronisch erzeugten PDF Dateien lassen sich schon viele Elemente finden und extrahieren. Text via PDFTOTEXT oder Bilder extrahieren mittels PDFIMAGES.

Aber Formeln stellen ein Problem dar. Wie zum Beispiel

$$\frac{\sqrt{x}}{2x^2}$$

endet z.B. mittels PDFTOTEXT wie folgt (Zeilen nummeriert):

1: -

2: √x

3: -------

4: 2

5: 2x

Ziel der Arbeit ist das automatische Finden mathematischer Ausdrücke (abgesetzt und Inline) inkl. der Ersetzung des Formeltextes durch eine MathML-basierte Vektorgrafik (SVG) und LaTeX-Quellcode als Alternativtext (für letzteres: Formelbild zu LaTeX, gibt es Vorarbeiten) innerhalb der PDF.

Aufgaben

- Einarbeitung in das Thema (PDF-Strukturen, Elementextraktion,...)
- Entwicklung eines Verfahrens zum Auffinden von Formeln
- Einsetzen vorhandener Formelerkennungsverfahren
- regelmäßige dokumentierte Evaluationen
- Rückführung der Erkennung in die PDF-Dateien (SVG inkl. Alternativtext anstelle der Grafik)
- Final: Entwicklung eines lauffähigen On-/Offline-Tools

Voraussetzungen

- Interesse an der Thematik
- Machine-Learning / Computervision Verfahren
- Freude am Einbringen eigener Ideen
- Programmierkenntnisse

SZS

Ansprechpartner für fachliche Fragen:
Dr. Thorsten Schwarz 0721/608-46888
thorsten.schwarz@kit.edu