Document Structure Extraction

Extrahierung von Text und Struktur eines PDF Dokuments

Research Studios Austria
DSc.FS:CADUL

```
<g id="g224">
  <g clip-path="url(#clipPath230)" id="g226">
    <text id="text234"
          style="font-variant:normal;font-weight:normal;font-size:11.03999996px;
                 font-family:ArialMT;-inkscape-font-specification:ArialMT;writing-mode:lr-tb;
                 fill:#000000;fill-opacity:1;fill-rule:nonzero;stroke:none"
          transform="matrix(1,0,0,-1,70.824,469.34)">
      <tspan
        id="tspan232"
        y="0"
        x="0 6.2265601 12.45312 17.97312 20.368799 23.48208 25.87776 31.154881 37.381439
           43.608002 46.489441 53.919361 59.914082 66.14064 72.135361">
        positiven Ende
      </tspan>
    </text>
  </g>
</g>
```

Abb. 1: Ausschnitt einer .svg Datei, welcher den Text „Positiven ende“ beschreibt.

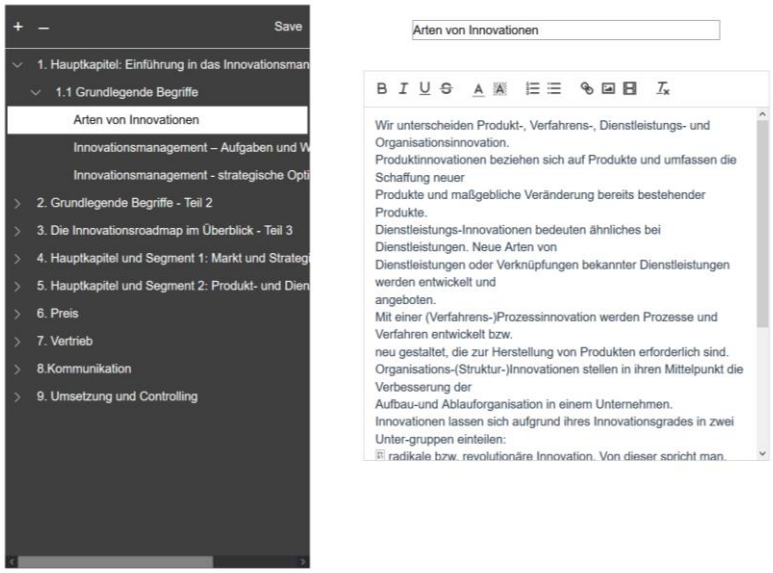

Abb. 2: UI zur Anzeige eines extrahierten Dokuments (Daten aus der Datenbank) in einer Baumstruktur

Hintergrund

Im Zuge der Feasibility Study: „Conversational AI für Domänen-übergreifenden Lerndialog“ wurde ein Prototyp entwickelt, mit dem es möglich ist Lerninhalte durch die Verwendung eines Chatbots abzufragen. Um so einfach wie möglich an Lerninhalte zu kommen, wurde eine Extraction Pipeline entwickelt, die aus bestehenden .pdf Dokumenten den Inhalt unter Beibehaltung der Struktur auslesen kann.

Methode

Die Extraction Pipeline besteht aus den folgenden Schritten:

1. Aufsplitterung des Dokuments in einzelne Seiten.
2. Umwandlung aller .pdf Dokumente (Seiten) in .svg Dateien mithilfe von Inkscape (Text wird als Text angezeigt und nicht als Vektorpfad. Der Inkscape Export Prozess garantiert nicht, dass sich ein(e) Wort/Zeile/Absatz im selben Element befinden.)
3. Analyse des Dokuments mithilfe eines XML Parsers (Layout des Dokuments kann aus Informationen wie den Koordinaten einzelner Tags und dem Font Style des Texts extrahiert werden. Es können ebenfalls Bilder extrahiert werden.)
4. Parallel zur Analyse der .svg Dateien wird der Text aus den .pdf Dateien mittels Apache Tika ausgelesen.
5. Zusammenführen der beiden Analysen (Apache Tika -> Text, SVG -> Struktur, Bilder)

Ziel

Ziel der Extraction Pipeline ist es bestehende Dokumente einlesen zu können und deren Inhalt/Struktur in eine relationale Datenbank abzulegen. Somit können die gewonnenen Daten sehr einfach in anderen Systemen verwendet werden.

Innovation

- Einfache Umwandlung existierender Dokumente zur universellen Weiterverarbeitung

Nutzen

- Basis für ein Intelligent Tutoring System (ITS) zur einfachen Übertragung von Lerninhalten aus bestehenden Dokumenten
- Umwandlung von Dokumenten in relationales Datenbank Schema unter Beibehaltung von Layout und Struktur

Demonstration

- The 21st International Conference on Information Integration and Web-based Applications Services (iiWAS2019)

Publikation

- Bernhard Göschlberger, Christoph Brandstetter; Conversational AI for Corporate e-Learning

RSAFG
Research Studio DSc

dsc.researchstudio.at | office@researchstudio.at