Building a better working world

Ernst & Young (Wirtschaftsprüfungsgesellschaft)

Die globale EY-Organisation ist einer der Marktführer in der Wirtschaftsprüfung, Steuerberatung, Transaktionsberatung und Managementberatung. Mit unserer Erfahrung, unserem Wissen und unseren Leistungen stärken wir weltweit das Vertrauen in die Wirtschaft und die Finanzmärkte. Dafür sind wir bestens gerüstet: mit hervorragend ausgebildeten Mitarbeitern, starken Teams, exzellenten Leistungen und einem sprichwörtlichen Kundenservice. Unser Ziel ist es, Dinge voranzubringen und entscheidend besser zu machen – für unsere Mitarbeiter, unsere Mandanten und die Gesellschaft, in der wir leben. Dafür steht unser weltweiter Anspruch „Building a better working world".

Titel des Projektes: PDF Daten Extraktor zur Erstellung normalisierter Datensätze

Projektauftrag:

Im Alltag kommt es häufig vor, dass Kollegen bestimmte Elemente aus PDF Dokumenten abtippen müssen. Die Dokumente haben dabei immer einen ähnlichen Aufbau. Ihr Team soll im Rahmen dieses Projektes aus einer Anzahl an Beispiel Dokumenten verschiedene Elemente automatisiert extrahieren und in einer sinnvollen Form aufbereiten.

- Bilder als JPG extrahieren
- Tabellen als CSV oder Excel
- Textbausteine strukturiert extrahieren
- Wiederkehrende Muster erkennen
- Normalisierte Datenstruktur zur Weiterverarbeitung

Projektziel:

- Erreichen des Projektauftrages
- Abbildung eines Auftraggeber - Auftragnehmer Szenarios
- Neue Einblicke und Lösungsansätze
- Networking

Anforderungen:

- Studium der Informatik oder Wirtschaftsinformatik
- Programmierkenntnisse
- Text, Tabellen, Bilder inkl. Formatierungen, etc. stabil aus PDF extrahieren, bspw. nach Word, Excel, etc.

Zusätzliche Informationen:

Der Extrakt von strukturierten Daten aus einem PDF Dokument ist eine häufige Herausforderung, insbesondere Tabellen, die zwar optisch als Tabelle erkennbar sind, aber technisch als Fließtext im PDF hinterlegt wurden stellen Folgesysteme vor erhebliche Hürden. Der Lösungsansatz kann sich gerne auch auf Konzepte wie OCR oder lernende Systeme erstrecken, sofern konventionelle Methoden kein ausreichend qualitatives Ergebnis liefern.

Anzahl der freien Plätze:

4-6

Einsatzort:

EY Office Köln (Börsenplatz 1, 50667 Köln)

Projektverantwortliche:

Alexander Simon: alexander.simon@de.ey.com
Marco Heßland: marco.hessland@de.ey.com
Julian Klein Pohlmann: julian.klein.pohlmann@de.ey.com